中文文本自动校对技术的研究

骆卫华　罗振声　宫小瑾
(清华大学中国语言文学系　北京　100084)
(lwh99 @mails. tsinghua. edu. cn)

摘　要　传统的自动校对技术多是基于字、词级的统计方法,有很多局限,通过讨论中文文本自动校对技术的设计思想与实现方法及中英文自动校对的异同,提出了词法、句法、语义多层次结合校对策略,从而能够检查以往无法查出的错误. 描述了自动校对系统的整体框架,并具体给出可操作的实现方法. 这些方法针对非受限领域的文本,为中文自动校对技术的发展提供了新的思路.

关键词　中文文本自动校对; *n* 元模型;主题相关度;语义共现矩阵

中图法分类号　TP391. 1

Study of Techniques of Automatic Proofreading for Chinese Texts

LUO Wei-Hua , LUO Zhen-Sheng , and GONG Xiao-Jin
(Department of Chinese & Literature , Tsinghua University , Beijing 100084*)*

Abstract　Traditional techniques of automatic proofreading mainly adopt stochastic methods on character or lexical level , and cause some limitation. By discussing the design thought and implementation methods of automatic proofreading for Chinese texts , a proofreading strategy of combining multi-levels of morphology , syntax and semantics is proposed , and consequently the errors that cannot be checked before are detected. The whole framework of the system of automatic proofreading is described and the operable implementation methods are introduced in detail. Designed for the text of unrestricted domains , these methods provide new thoughts for future development of proofreading for Chinese text.

Key words　automatic proofreading for Chinese texts; *n*-gram modal; degree of subject relativity; semantic co-concurrence matrix

1　校对技术的发展

中文信息处理技术的发展加快了出版节奏,同时也需要缩短校对周期,而传统的手工校对效率低、强度大、周期长. 因此,从此种角度来说,文本自动校对的实现已成为中文信息处理技术发展的关键所在.

英文自动校对的研究大约始于 20 世纪 60 年代. 20 世纪 80 年代初期,美国一些大学对英文单词词长和音节划分等规律进行了研究,对文本中的错误做了总结. 同时,一些学者提出了拼写检查和语法检查的方法,如三元语法相似度评测方法,对英文错误进行自动拼写改正. 英文文本的自动校对基本以词的校对为核心,在"非词错误"和"真词错误"两个层次上进行[1]. 非词错误指文本中被词边界分隔出的字串而不是词典中的词条. 真词错误指虽然字串是词典中的词,但它与上下文搭配不当. 其查错和纠错方法也相应地分为孤立词和上下文相关的两种方法. 英文文本中非词错误的比例比较大,所以较容易做出实用的英文校对系统.

20 世纪 90 年代初,国内的研究人员开始探索用电脑来进行中文自动校对的可行性. 初期的主要设计思想还是从国外引入的. 但由于汉语本身的特点,英文的研究方法不能完全适用. 目前中文文本自动校对的研究总体上还处于刚起步的阶段,采用

收稿日期:2002-02-27;**修回日期**:2003-07-30

的方法多是字、词级别上的统计方法,使用的模型较简单,利用的语言学知识也不丰富. 对一个有经验的校对人员来说,自身丰富的背景知识、专业知识、语言知识、经验知识等对最后的校对结果起着非常重要的作用. 而对计算机来说,要实现人所具有的这些能力在目前还无法做到,所以必须另辟蹊径,借助统计等手段来揭示语言的内部规律,并结合规则、词典等工具来提高校对的性能指标.

2　汉语文本常见错误分类

根据错误的文字或符号,文本错误可分成文字错、标点错、数字错和其他字符错几大类. 其中文字错占错误的大多数,也是自动校对研究的重点. 根据文字错的表现形式可分为以下几种.

(1) 代换错,又可以分成:

①一对一的对称代换错

例 1. "清华园原是康熙时熙春园的一部分,建圆时间与附近的圆明园相同"中的"建园"错成了"建圆".

②一对多、多对一、多对多的非对称代换错

例 2. "全校师生一起倾注建校 90 周年"中的"庆祝"错成了"倾注".

(2) 非代换错,包括:

①缺字错

例 3. "项目应该提交上级主管部审批"中"主管部门"漏掉了"门"字.

②加字错

例 4. "至之 1908 年,清朝与美国协议用庚子赔款的部分退款兴学育才"中"至"后面多了个"之"字.

③换位错

例 5. "这个密秘只有我们俩个知道"中"秘密"错位成了"密秘".

表 1 是在对报社 305 个校样中的错例进行分析之后,得到的错误比例[2]:

表 1　错误类型及其分布

错误类型		比例/ %	
文字错	对称代换错	89	64.9
	缺字错		17.7
	多字错		9.9
	其他错		7.5
标点符号错		9	
英文拼写错		1	
其他错		1	

3　中英文自动校对的比较及面临的困难

相对英语而言,中文文本自动校对面临的困难主要有:

(1) 汉语输入是人工编码的间接输入,录入时的输入法不同,其错误的类型也大不一样;

(2) 汉语的词之间没有分隔标志,在进行任何词和词以上层级的处理时,都必须先分词,分词结果直接影响到校对的查错率、召回率,但其本身尚有歧义切分、新词等难题尚未解决;

(3) 汉语的词类没有形态上的标志,而且和句法成分之间并没有简单的映射关系,再加上兼类、句法成分省略等的干扰,使得汉语的句法分析困难重重;

(4) 随着与国外文化、科技交流的逐渐增多,汉语和英语混用的情形变得很普遍,在目前的技术条件下,只能分开处理,从而导致句子成分的缺失,这也增加了文本校对的难度.

上述问题,有些可以具体问题具体解决. 比如分析发现很多错字与录入人员当时使用的输入法有很大关系[3],如五笔字型输入法造成的形近字错、拼音输入法造成的音近字错等,通过总结这些输入法的编码规则,既可以发现错误规律,又能据此总结出常见错误的混淆集,在系统中往往能产生其他方法难以达到的效果.

4　文本自动校对的总体策略与面向问题的解决方案

总的来说,目前文本校对的理论和技术都不太成熟,但比较一致的看法是,自动查错和纠错应该在词法、语法和语义 3 个层次上进行. 实验系统采用了如下策略:①规则与语料库统计相结合;②词法、语法、语义多层次结合查错、纠错;③机器自动查错与人工确认纠错相结合;④面向查错的"粗分析"方法,具体问题具体解决. 分析采用逐句(即末尾为".","!"和"?"等结束符的字符串)分析的方式进行. 实施步骤主要分为 4 个层次. 首先对要分析的句子进行分词,在分词过程中如果遇到不能成词的单字或异常的单字,则认为有错. 然后,对已分词的句子进行词性标注并同时进行词性邻接的检查. 第 3 步,对已标注的句子进行句子成分分析,方法是先用上下文无关文法(自顶向下)解析句子各元素,

后用自底向上的短语匹配进行分析,并根据分析结果检查错误. 最后,进行局部语义分析,并用它来检查一部分错误.

为客观地评价查错与纠错系统的性能,引入以下一些概念:

召回率(r) = 查出预校正文本中真正错误的总数/ 预校正文本中错误的总数;

精确率(a) = 查出预校正文本中真正错误的总数/ 查出预校正文本中错误的总数;

误报率(e) = 所有的虚假错误总数/ 查出预校正文本中出错的总数.

根据现有系统已公开的数据,可发现一般都是召回率越高,精确率就越低,而误报率越高. 比如清华大学中国语言文学系的实验系统召回率为 89 %,精确率为 40 %[4],东北大学的 HMCTC 系统的召回率为 62 %,精确率为 38 %. 对于一个全自动的校对系统,其误报率必须小于 50 %,否则这样的系统不仅没有减少错误,反而增加了错误.

5 中文文本自动校对系统的实现方法

5.1 系统框架

实验系统在使用比较成熟的自动分词、*N* 元邻接矩阵等方法的基础上,结合语法、语义研究的新成果,使自动校对在句法和语义级上取得进展. 系统的粗框架如图 1 所示:

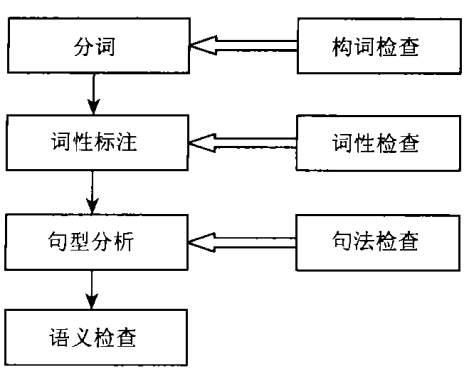

图 1 自动校对系统流程

框架中每一个模块都实现了一个层级上的分析,并把分析结果传递给下一级. 它们的功能是相对完整和独立的,每一级算法的改进都将带来整个系统性能的提高. 系统的主要策略如下所述.

5.2 词法校对

词法校对以自动分词为基础,结合统计和规则的方法,校对出不符合现代汉语词法规律的词级错误.

本系统采用的是正向最大匹配和逆向最大匹配相结合的双向匹配法. 校对系统中的分词是为校对服务的,它无需解决某些与校对无关的分词难题,比如"新研究生宿舍"这样的歧义问题. 另一方面,校对中的分词还肩负着查错以至纠错的任务. 在这个阶段中最有可能查出的错显然是不成词的字,但是它的效率并不高. 主要是因为: ①汉字的单字成词能力很强; ②现有的任何系统都无法识别出所有新词,所以就会有一定的误报. 比较基本的做法是先进行最大匹配的分词,然后根据结果进行检测,切分碎片就是可能的新词,考察其上下文环境,看是否满足相应的激活条件,如满足就执行进一步的处理.

(1) 未登录词处理

多数新词在文本中不止出现一次,通过标记新词出现次数,能认出大部分新词来. 首先设定判断新词的激活条件: ①不能成词的字; ②频率很低的单字词; ③频率很低的单字词的词串. 对一个不能成词的字,记下它前面已切出的词,作为该字的上文,然后跳过该字继续切分. 如成功,则记下该字段的首词作为该字的下文,如不成功,则记下该字后面第 1 个字作为该字的下文. 对一些很少出现的单字词,系统直接记下其前后字作为上下文. 对一串相邻的单字词,先对词串大于 3 的串进行分割,直到每个词串的词的个数都不大于 3,对 3 个词的词串,定义中间词为中心词,则前后词就是它的上下文,两个词的词串,则定义后一个词为中心词,同样记下它的上下文环境. 利用筛选算法,选择一个与该字结合最佳的上文或下文,从而结合成一个可能的新词. 筛选算法的核心思想是从上文和下文中各选一个出现最多的词进行比较,选择出现次数显著大且高于阈值的一方. 例如在一篇文章中,"核"的下文出现最多的是"酸",共 15 次,上文出现最多是"细胞",共 4 次,"酸"显著多于"细胞",因此认为"核"与"酸"结合成"核酸",然后把所有的"酸"记录删除,进行下一轮比较. 这种方法无论从直观上还是在实际应用中都是有效的.

(2) 歧义处理

歧义问题是分词的另一个难点,我们不求彻底解决这个问题,但也不能忽视. 歧义处理的总体策略是规则与统计方法相结合,在分词的后处理中使用统计的方法,计算歧义字段不同切分的词语共现概率,取大的作为正确的切分结果. 设计一个评价函数,对切分的结果进行综合评价. 考察汉语的特点,可以发现多数情况下修饰语在中心语的"前面",

因而借助于一条无需任何语言知识的“归右原则”就能实现正确的切分,这条原则规定交叉歧义字段优先与其右边的字段成词. 一般而言,“归右原则”可以使分词的精度上升到 99 %以上[2].

(3) 模糊匹配

对于四字或四字以上的长词,考虑用模糊匹配的方法进行处理. 四字以上的长词在汉语中一般都属于固定搭配,不允许更改其中的字和词序. 如果模糊匹配成功,则认为该词发生了错误,同时还可完成纠错工作.

(4) 人名地名的识别

建立姓氏名字用字频率表、称谓表、指界动词,如“说”、“认为”等信息,考虑姓名在句中出现的位置和分布情况是一条行之有效的途径. 地名的识别比人名要复杂,除了建议常用地名词典及利用一些标志性词,如“市”、“村”、“湖”等,还没有出现有效的方法进行判别.

使用统计方法就是从语料库中计算字频、二元字邻接矩阵、二元(三元)词邻接矩阵、二元词性邻接矩阵. 统计方法不依赖于知识,因而应用起来比较方便. 统计应用一般是在若干个候选方案中进行选择时计算不同方案之间的优劣. 比较优劣的方法是:

设文本中的句子为 $S = C_1 C_2 \cdots C_n$,经分词后 $S = W_1 W_2 \cdots W_m$,其中,C_i 为第 i 个字,W_i 为第 i 个词. 对于字段 $C_i \cdots C_j$ 计算以下数据:

字段平均概率 $F_1 = (ZF(C_i) + \cdots + ZF(C_j))/(j - i + 1)$;

字段转移概率 $F_2 = P(C_i | C_{i+1}) \times \cdots \times P(C_{j-1} | C_j)$,其中,$P(C_k | C_{k+1}) = R(C_k | C_{k+1})/(R(C_k) \times R(C_{k+1}))$,$R(C_k | C_{k+1})$ 为二元字字同现次数,$R(C_k)$ 为字频.

对于词段 $W_i \cdots W_j$ 计算如下数据:

词间字转移概率 $F_3 = P(I_i | I_{i+1}) \times \cdots \times P(I_{j-1} | I_j)$,其中,$P(I_k | I_{k+1}) = R(I_k | I_{k+1})/(R(I_k) \times R(I_{k+1}))$,$R(I_k | I_{k+1})$ 为二元词间字同现次数,$R(I_k)$ 为词间字频.

词性转移概率 $F_4 = P(T_i | T_{i+1}) \times \cdots \times P(T_{j-1} | T_j)$,其中,$P(T_k | T_{k+1}) = R(T_k | T_{k+1})/(R(T_k) \times R(T_{k+1}))$,$R(T_k | T_{k+1})$ 为二元词性同现次数,$R(C_k)$ 为词性在统计语料中出现的次数.

对于一个字段计算 F_1 和 F_2,对于一个词段计算 F_3 和 F_4,如果某个值低于正常范围,则怀疑此字段或词段有错,并进一步分析与判断. 这一步最主要的问题是如何确定“正常范围”,目前的方法是对每个结果设定阈值,如果结果低于阈值,则认为其有错. 在实际的系统中,必须根据实际效果不断调整阈值.

在系统中还使用了词性 N 元模型查错,由于一个词的词性可能不止一种,因此必须额外引入自动词性标注过程. 同时,在标注的过程中,还可以对邻接词性的搭配进行检查,以发现搭配错误. 对一给定的词串 $W = W_1 W_2 \cdots W_n$,词性标注的目标是要找出一个标记串 $T = T_1 T_2 \cdots T_n$,使得 $\prod [P(T_i | T_{i-1}) P(W_i | T_i)]$ 的值最大,寻找最大值的过程也是标记选择的过程. 但目前的自动词性标注技术采用的也是 N 元模型,其基本任务是寻找一条标记间接续强度最大的标记路径,这时,再使用与标注过程相同的依据进行查错就陷入了“鸡生蛋,蛋生鸡”的怪圈,无疑会掩盖掉很多错误. 因此必须进行适当的后处理.

词级查错中另一个值得关注的问题就是数据稀疏问题,这也是自然语言处理中难以避免的问题. 实际使用中要通过平滑或聚类等方法来解决.

5.3　句法校对

句法分析是通过检查句子语法成分的搭配关系来判断是否存在句法上的错误,它是语义分析的基础. 但是校对系统中的句法分析跟一个独立的句法分析系统不同,在很多时候并不需要将句子分析出一棵语法树,只能算做是一个句法的粗分析. 在句法分析的过程中结合语义分析,可以尽早发现句子中的错误部分,也可以补偿句法分析本身的缺陷导致的漏查.

5.4　语义校对

目前还不可能让计算机像人一样去理解一句话的意思. 一个词法、语法合格但语义不通的句子在缺少语义分析的校对系统看来就是正确的. 同时,经过上面的层层过滤,可以认为沉淀到这一层的错误都是相对“高级”的错误,需要借助更高层级的语言知识才能处理.

一种方案是仿效词级校对,使用词义的 N 元邻接矩阵进行查错(N 取 2 或 3). 语义属性的抽取依赖于所用的词典,实验系统采用的是《知网》词典. 同样定义语义平均概率和语义转移概率两个评价参数,把计算结果与阈值相比较,如果显著低于阈值,则认为该句有错. 但这样一来又无法知道出错的是哪个词. 一种简单的方案是使用相似度来处理. 相

似性计算可以采用以词为匹配单位的算法,它不需要语法、语义方面的信息,实现起来比较简单. 但它需要有跟错句做参照的正确语句,相对于近乎无限的自然语言的句子,语料库中收集的句子是非常有限的,如果找不到参照系,这种方案也就失效了.

词义的 N 元邻接矩阵的缺陷显而易见:首先,对词典中的 6 万多词构建二元邻接矩阵,工作量相当巨大,面临的数据稀疏问题更严重. 其次,它只能检查相邻词的语义搭配关系,如果两个相隔较远的词搭配不当,N 元矩阵就无能为力了(除非 N 取值很大,但将会造成算法的极度复杂而无法实现). 而结合主题相关度和语义共现矩阵的方案将会克服这些缺陷,并为自动查错系统的语义检查开辟一条新的途径.

众所周知,一篇文章都会有一个主题,这个主题将会在文中通过词语和句子不断得到重现. 如果设想某篇文章从头到尾就是一句话,那就把篇章映射到了句子上. 由此可以假设,一句话也有一个主题并由这句话的词来体现. 当然,视句子长短而定,这个主题可能重复一次、两次,乃至若干次. 当句子中出现了语义搭配不当的问题,就可以认为其中的一个词偏离了主题. 如何把这种"偏离"量化成可操作的算法,是这种方案实施的关键.

同词性的兼类一样,一个词的词义也可能有好多种,并且分属不同的领域. 参照自动词性标注过程,设法找出一条语义接续强度最大的路径. 设一个句子 $S = W_1 W_2 \cdots W_n$,目标是找出一个语义串 $M_1 M_2 \cdots M_n$,使得 $\prod [P(M_i|M_{i-1}) P(W_i|M_i)]$ 值最大. 只有定出惟一的语义,才能确定话语的主题及其相互关系.

根据知网的定义,每个词的主要特征实际上是该词所属的领域,对于范畴较大的领域,其附加特征进一步缩小了该领域[5]. 比如,"银河"属于"天体","银行"属于"处所",这个范畴就比较宽泛了,因此后面又加了一个限定性特征"钱财". 在确定了每个词的语义之后,就可以提取一句话的主题了. 一般的做法是,首先提取每个词的首特征,对于有附加特征的则取附加特征,对词 W_i 的领域 F_i 和 W_j 所属的领域 F_j 进行比较,如果不相同,则分别对 F_i 和 F_j 上溯,寻找它们的最小公共域 F_x,并记录下来,然后用最小公共域分别替代原来标记的领域. 比如"天体"和"大地"都是域,它们同属于"天然物". 如果上溯到最顶端都没有公共域,就把原始域记录下来. 最后把所有这些域进行累加,取累加值最高的域作为本句的主题. 考虑到句子的成分不同,重要性也不同,因此对各个部分赋予一个权值,并根据实际应用的效果进行调整. 直观地说,句子的主语中心词应该被赋予更高的权值,而其他修饰性成分相对来说没那么重要,其权值应该低一些. 按照累加值从高到低的词序对提取出的域特征及其对应的词进行排序,然后考察这些词与主题的偏离程度. 知网中每个特征都有一个代码,这些代码组成了一个树形图,下一级的结点号是上一级的结点号再加一个与同级结点相区别的代号组成的. 比如"钱财"的代码是"N.1.1.1.2.2.9","货币"的代码是"N.1.1.1.2.2.9.4","货币"就是"钱财"的一个子结点. 我们认为所属领域同主题的代码倒数第 2 位不同的词就是"可疑"的,比如主题的代码是"N.1.1.3",其中一个词的领域代码是"N.1.2.1",那么这个词就是"可疑"的,需要进一步分析. 由于语义知识只能推导到这一步,所以下面的工作还是需要借助统计手段,通过在语料库中建立一个语义共现矩阵,可以发现哪些语义搭配是罕见的. 在标记出"可疑"词之后,在语义共现矩阵中查找其共现概率,如果低于阈值,就认为这个词是错的,从而完成语义查错.

比较以上两种方法,可以发现它们各有所长,较短的句子主题不太明显,甚至可能因为话题词是错的而误判了本句的主题,因此比较适于用邻接矩阵来查错. 对于长句,邻接矩阵往往鞭长莫及,通过查找建立话语主题的方法,就可能找出以前无法发现的错误.

6 结束语

本文详细介绍了中文文本自动校对的设计思想,提出一些可操作的实现方法. 在词法校对方面,针对未登录词、歧义、固定搭配识别等具体问题给出了新的解决方法. 对于目前尚无太多研究的句法语义校对,提出主题相关度和语义共现矩阵相结合的处理策略,从而能够发现单纯用词法检查无法校对出的错误,提高了系统的召回率.

但是对查错后的纠错处理,本文并未提出好的可行性建议,这是因为除了混淆集之外,目前对如何提出纠错候选词的研究没有太多进展. 对句法和语义分析这一块,综合已有成果提出了一些想法,但还有待在实验系统检验,算法复杂度有可能超出预期.

显而易见,无论是句法还是语义校对都要建立在一部好词典的基础上,现有的词典虽然有了一定的规模,但无论在广度还是精度上还都有待完善. 这些问题都需要在以后的研究工作中解决.

参 考 文 献

1 K Kukich. Techniques for automatically correcting words in text. ACM Computing Surveys, 1992, 24(4): 378～431

2 慕勇. 清华语料库的研制与汉语文本自动校对的研究[硕士论文]. 清华大学计算机科学与技术系, 北京, 1995
(Mu Yong. The building of Tsinghua corpus and the study of automatic proofreading of Chinese texts[Master dissertation](in Chinese). Department of CS & T, Tsinghua University, Beijing, 1995)

3 张仰森, 丁冰青. 中文文本自动校对技术的现状及展望. 中文信息学报, 1998, 12(3): 50～56
(Zhang Yangsen, Ding Bingqing. Present condition and prospect of Chinese text automatic proofreading technology. Journal of Chinese Information Processing(in Chinese), 1998, 12(3): 50～56)

4 黄晓宏. 汉语文本自动查错和确认纠错系统的研究[硕士论文]. 清华大学计算机科学与技术系, 北京, 1996
(Huang Xiaohong. An approach to detect and correct errors automatically in Chinese texts[Master dissertation](in Chinese). Department of CS & T, Tsinghua University, Beijing, 1996)

5 知网及其说明文档. 2001. http://www.keenage.com
(Hownet and its specifications. 2001. http://www.keenage.com)

骆卫华　男,1977 年生,硕士研究生,主要研究方向为中文自动校对技术、信息提取等.

罗振声　男,1938 年生,教授,硕士生导师,主要研究方向为自然语言处理技术.

宫小谨　女,1979 年生,硕士研究生,主要研究方向为中文自动校对技术、句法分析等.