文章编号：1003-0077（2014）00-0000-00

面向政治新闻领域的中文文本校对方法研究[1]

张仰森 唐安杰 张泽伟

（北京信息科技大学智能信息处理研究所，北京，100192）

摘要：政治新闻领域内文本错误多为语义级错误。在研究新闻领域文本政治性差错的语言表述特征的基础上，分析了报刊新闻中政治性差错的表现类型，构建了面向各类错误侦测的词库和知识库。通过研究政治新闻文本的语言学特征，提出了一个政治性差错文本错误侦测规则的一般形式化模型，采用统计与规则相结合的策略实现政治新闻领域文本的语义校对。实验结果显示，该方法的召回率为65.5%，精确率为80.5%，具有较好的应用前景。

关键词：政治新闻；文本校对；查错模型

中图分类号：TP391.1　　**文献标识码：**A

Research on the Method of Chinese Text Proofreading Oriented to Political News Field

ZHANG Yangsen，TANG Anjie， ZHANG Zewei

（Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing, 100192）

Abstract: Most of the errors in the political news field are semantic errors. On the basis of researching language expression characteristics of news field text political error, we analyzed the political error performance type in newspapers news, and built some relevant knowledge bases for political error detection. According to the research on linguistic features of political news，the formalized model of detecting political errors is presented. The strategy based on the combination of rules and Statistics is used to proofread semantic errors of the political news field. The results show that the recall rate is 65.5% and accuracy rate is 80.5%, which reflects the good application prospect.

Keywords: Political news; Text proofreading; error detecting model

1 引言

当今网络传媒快速发展，报纸种类也越来越多，竞争异常激烈，各类差错也如影随形。有的报纸违纪违规，发生导向错误，有的采访不深入，出现新闻失真或虚假新闻，有的在涉及港台澳以及国家主权方面出现错误，这些错误都是影响比较大的，甚至影响国家的稳定[1]。因此，研究政治新闻领域的文本校对技术意义非常重大。然而，新闻中的错误除了一些印刷错误外，很多错误可能是影响舆论导向的政治性错误，是潜在的语义级错误，采用通常的文本校对方法，很难发现这些错误，而是要检查语句中所表达的语义和语用是否违背了某种标准，例如，报刊、网络文章中出现的一些关于台湾问题的不正确表述等，难度是相当大的。但政治错误对报刊杂志的影响是很大的，是编辑部校对的重中之重。采用人工校对，劳动强度大，成本高，还由于校对人员的责任心或视觉疲劳等问题，会漏掉许多错误；采用计算机自动校对技术侦测政治性错误，由于难度大，目前相关的研究比较少。王燚[2]通过将句子和短语结构转为一阶谓词逻辑表达式，匹配标准库中的标准，实现台湾问题的语义匹配。但是局限于逻辑表达式的规模，仅实现了台湾问题部分检查，并且推理效率比较低。本文利用新闻领域政治性文本中的语言学特征和统计特征，细化文本中政治性差错的错误类型，提取相

[*] 基金项目：国家自然科学基金项目：（编号：61070119，61370139）、北京市属高等学校创新团队建设与教师职业发展计划项目(IDHT20130519)和北京市教委专项基金：科研基地-科技创新平台-面向内容理解的智能信息处理研究平台(PXM2012-014224-000020)

关知识库，制定政治性差错侦测规则库，并提出了查错规则的一般形式化模型,采用统计与规则相结合的策略对文本进行多级查错和分类查错，以实现政治新闻领域文本的语义校对。

2 新闻领域文本政治性错误类型分析

政治性差错从表现形式上来看主要分两类：一是直接陈述出来的，二是通过字里行间表现出来的[3]。对于字里行间表现出来的隐性错误，利用计算机实现自动错误侦测难度很大，而对于直接表现出来的政治性错误，通过查阅文献[4]以及对相关中央文件和网络文本的统计分析,发现以下四类错误是政治性新闻领域出现频率较高的错误类型：

（1）政治性或政策性错误。这类错误主要涉及意识形态领域的政治倾向错误、损害国家利益的错误、违反民族政策方面的错误言论、领土主权及港台澳问题上违反国家政策法规的错误等等。由于政治性或政策性错误一般属于语义级错误，只有具有高度政治敏感性的人才能发现其错误，一般的自动校对只是从文本结构中发现错误，因此，难度比较大。涉及国家领土、主权和港台澳问题的错误，由于有中央文件《涉及港澳台用语规范 34 条》和《新华社新闻报道中的禁用词》[5]作参考，本文将对涉及港台澳问题的错误进行校对。例如，在有的报纸中出现“中港合资”、“中台合资”的情况。再比如，2004 年 11 月，某报在报道孙楠的一篇文章中，竟把香港一唱片公司说成是“境外”的唱片公司。

（2）领导人姓名错误。即新闻文本中涉及的领导人的姓名出现错字、别字、多字或少字。例如，2010 年 12 月 30 日出版的《人民日报》第 4 版的文章标题将国务院总理温家宝姓名错印成“温家室”，2005 年 3 月 15 日，有一标题为“消费者的烦恼与期盼”的消息，文中将“温家宝总理”错成“温家保总理”。幸亏这一领导人名的重大差错在校对环节被堵住，否则对媒体将产生极其不良的社会影响。此类错误侦测需借助句子中的特征词，如相应职务等，若无相关的特征词则很难校对。

（3）领导人顺序错误。即文本中出现的国家领导人姓名次序不符合领导人的职务排位顺序。例如李克强、习近平、李源潮等出席了本次会议。句中“习近平”与“李克强”次序颠倒。

（4）姓名-职务对应错误。即文本中涉及的领导人姓名与其职务不符合规定。错误形式主要包括：①领导人姓名正确但是职务搭配不正确，如“国务院总理习近平”中习近平对应的职务应为中国中央总书记、国家主席、中央军委主席。②领导人姓名正确、职务搭配也正确，但是职务顺序不正确，如“国家主席、中共中央总书记、中央军委主席习近平”中,国家主席应该排在中共中央总书记之后。③领导人姓名和职务均正确但出现了重复，如“中共中央总书记习近平总书记”中两个总书记出现了重复。④领导人姓名错误但是职务正确，如“全国政协主席俞正生”⑤领导人姓名正确但是职务错误，如“中国中央总书记习近平同志”⑥领导人姓名和职务均出现错误，如“中国中央总书记刁近平”。

（5）输入过程疏忽造成多字、漏字或别字而引起的政治错误。这类错误有时会出现在新闻的正文里或标题中，如几年前某报在美国前总统克林顿访华之际，将“克林顿访华”错成“克林顿反华”[1]。这是一个严重的政治差错，会引起读者的不解或外交风波。对于多字、漏字等错误，其校对方法和一般文本的校对方法类似。

在真实文本中，以上各类型的错误有时会出现一定的交叉重叠，如在多位领导人姓名并列时，若出现某个领导人或几个领导人的姓名错误，则可能会造成领导人顺序错误。

3 面向政治新闻校对的相关知识库构建

通用的中文文本自动校对系统的研究重心为查错和纠错算法，其构造的查错知识库也是面向所有领域的通用型知识库[6-8]，这就使得当针对特定领域文本进行自动校对时，由于通用知识库中包含有较少该领域的专业词语或知识，导致查错准确率和召回率下降。因此，构

建面向政治新闻领域的专业词库和相应的查错知识库，对于提高政治新闻文本自动校对系统的性能是非常必要的。

3.1 涉及主权、领土完整及港台澳问题的“引号词”QTLIB库的构建

《新华社新闻报道中的禁用词（第一批）》[5]第四部分对涉及我领土、主权和港台澳的禁用词列出了的13条规定。例如，（1）在涉及港台澳时，不能将其称为“国家”，尤其是多个国家和地区名称连用时，一定不能漏写“（国家）和地区”字样；（2）在涉及台湾当局“政权”系统和其他机构的名称，无法回避时应加引号，如台湾“立法院”“行政院”“监察院”等；（3）在涉及我国领土钓鱼岛时，不能将其称为“尖阁列岛”……。我们将涉及以上新闻中的带引号的敏感词称为“引号词”，并引入如下定义:

定义1：“引号词”是指在新闻领域政治性文本表达中需要加注双引号的字符串。

算法1.“引号词”库QTLIB的构建方法如下：

Step1. 提取2000年《人民日报》语料中加双引号的词；

Step2. 根据《中华人民共和国国家标准标点符号用法》中的引号用法规定，对引号词或短语进行筛选，除去引用他人话语和表示着重论述的引号内的词语或语句，保留具有特殊含义的词语和短语，形成候选集；

Step3. 计算候选集中各词语的政治敏感度，词语A的政治敏感度定义如下：

$$\alpha_A = \frac{\text{语料中词语A带引号的次数}}{\text{语料中词语A出现总次数}} \quad (1)$$

Step4. 如果$\alpha_A \geqq 90\%$，则将词语A加入引号词库QTLIB；

Step5. 依据《新华社新闻报道中的禁用词》，将一些涉及主权、领土完整和港台澳的引号词加入引号词库QTLIB；

Step6. 结束.

通过Step3和Step4的筛选，将会将那些只在少数特定语境下才加双引号的敏感词滤除掉，以提高查错的准确率，Step5则进一步滤掉了非政治性的加引号词。引号词库的格式见表1。

3.2 领导人顺序和姓名-职务知识库的构建

领导人顺序库以及姓名-职务库主要是为查找领导人顺序错误、姓名错误以及姓名-职务对应错误服务的。

3.2.1 领导人顺序知识库 LSQLIB 的构建

领导人顺序库的构建，根据国家级领导人的职务排位顺序，罗列出各领导姓名及其次序。对相同职务的领导人按其姓氏笔画排列，但七名中共中央政治局常务委员会委员的排列顺序由国家规定，不按照姓氏笔画排列。拥有多个职务的领导人，按照其最高职位在职位中的顺序进行排列。已卸任的国家领导人根据其卸任时间和卸任之前的职务进行综合排序，暂将已卸任的国家领导人排在现任国家领导人之后。领导人顺序库的格式见表 1。

3.2.2 领导人姓名-职务知识库 LNDLIB 的构建

该库主要为查找领导人姓名错误和姓名-职务对应错误服务的。通过对 2000 年《人民日报》标注语料中每位领导人的姓名与职务对应问题进行了统计分析，发现出现在领导人姓名前面的通常是其所担任职务。例如，在“江泽民”前面，通常会出现“中共中央（中国共产党中央委员会）总书记、国家主席（中国国家主席、中华人民共和国主席）和中央军委（中央军事委员会）主席”，且新闻稿中会根据所参加的不同活动，在不同的上下文中以三种形式出现，即（1）三个职务都出现，（2）出现两个职务，（3）出现一个职务。第（1）种情况一般在整篇报道的首段，或每段的前两句，第（2）种情况多出现在句首，第（3）种情况大部分出现在句首，少部分出现在句中。根据以上规律，制定相应的查错规则，有针对性，可

节省空间，提高效率。我们将出现在领导人姓名之前的职务称为**前职务项**。

统计发现，出现在领导人后面的词，可能是职务词，也可能是其它词。例如，在“江泽民”后出现较多的词有：同志（2051 次）、主席（1694 次）、总书记（1048 次）、说（370 次）、在（268 次）、今天（181 次）、指出（167 次）。其中“同志”、“主席”和“总书记”均为名词词性（/n），我们取名词词性的词“同志”、“主席”、“总书记”构成姓名-规则库的**后称谓项**。

我们构建了 13 个词库和知识库，由于篇幅限制，这里不一一介绍。每个知识库都具有良好的可扩展性，各知识库的组织形式如表 1 所示：

表 1 词库和知识库

库名	存储格式	说明
常见错误词库		
引号词库	每行一个词	
领导人顺序知识库	习近平+1	“+”后为领导人次序
姓名-职务知识库	习近平：中共中央总书记、国家主席、中央军委主席？总书记、主席、同志	“:”和“？”之间是前职务，“？”后是后称谓
国外重要政要库	姓名+职务	
…	…	…

4 面向政治新闻领域的差错侦测算法与实现

4.1 政治性差错侦测规则库构建

政治性差错的侦测主要涉及第2节提出的5类错误，对于第1类至第4类的港台澳问题和领导人姓名、顺序、职务错误，依据所构建的词库或知识库，以规则算法实现错误侦测，而第5类错误则利用统计语言模型实现错误侦测。由于篇幅限制，这里只给出“引号词”错误侦测算法和领导人顺序错误的侦测算法。

4.1.1 涉及港台澳问题的相关错误侦测

涉及港台澳的问题，主要依据文献[5]中新华社的相关规定设计错误侦测算法如下：

算法2. 港台澳相关问题的错误侦测算法

Step1.利用文本分类算法判定文本是否为涉港台澳文本，是则转Step2，否则转Step5；

Step2.提取含香港、台湾、澳门等词的语句，检查句末是否有“国家和地区”，是则转Step3，否则，将该语句标红，转Step3；

Step3.对每个 $W_i \epsilon$QTLIB，检测W_i在新闻文本中是否出现，若出现并被双引号标注，则取下一个词，否则，将该词标红，转Step3循环，直至QTLIB中所有词检查完毕，转Step4；

Step4.检查文本中是否出现“文书验证”、“司法协助”、“引渡”、“两岸三地”、“两岸四地”等被双引号括起来的词，若出现，则对这些词标红，转Step5；

Step5.结束.

4.1.2 领导人顺序错误侦测

领导人顺序错误检测可能会有两种情况出现：一种是介绍多位领导人同时出席各类党政会议或经济、文化、体育等活动的新闻稿，如“出席会议的领导同志还有：王刚、王兆国、王岐山、回良玉、刘淇、刘云山、刘延东、……”；另一种是新闻报道中存在一位领导人向另一位领导人转达问候的文章，如 “杨洁篪首先转达习近平主席和李克强总理对普京总统的亲切问候”。第一种情况可直接利用稿件中的领导人姓名排序与 LSQLIB 库中的顺序进行比较，第二种情况则需要考虑“转达”之后的领导人的排序。为此，引入“传递性动词”的定义。

定义2:“传递性动词”是指主语作为中间人而进行传递或传达动作的词语。如 “转达”、

“传达”、“表示”、“说”、“指出”等。

传递性动词之后的领导人顺序一般按领导人顺序库LSQLIB中的规则进行排序检查，前面作为主语的领导人顺序不需要进行检查。领导人顺序检查算法如下：

算法3. 领导人顺序检查算法

Step1. 读入下一个含有领导人姓名的句子；

Step2. 检查该语句中是否有“传递性动词”，若有，按顺序从左到右提取传递性动词后的领导人姓名，否则，直接按顺序从左到右提取语句中领导人的姓名，存于一数组中；

Step3. 依据对数组中每个领导人在LSQLIB查找其次序编号，记入数组；

Step4. 比较各位领导人的编号，如果序号大小正序递增，则转Step1，否则，对出现反序的领导人姓名标红，转Step5；

Step5. 若文本检查未结束，转Step1，否则，转Step6；

Step6. 结束.

由于政治领域文本本身具有较高的敏感度，相关的错误语料相对较少，我们通过国家各级政府机关相关的指导性文件和网络资源进行规则的分析和制定。根据不同的错误类型在报刊中出现的频率和现实中产生的影响，对不同的类别的政治性差错制定了不同粒度的错误推理规则，共有78条推理规则，由于篇幅所限，其它的错误推理算法就不在这里列出了。

4.2 面向政治新闻领域的文本分词优化

我们调用了 ICTCLAS 词法分析系统并对其进行了一定的优化，将表称职务的几个词合成一个词，例如“国务院/nt 副/b 总理/n”变为“国务院副总理/pos”，同时，变更了一些敏感词的词性标注，如原有系统中国家和地区的词性标注都为 ns，优化后国家的词性标注变为 ct。为此我们定义了一个新词库 UserWord，共包含 462 个词条，每条知识是一个二元组，用（W,P）表示，W 表新词，P 代表词性，词性标注类别主要有表 2 所列的几种：

表 2. 新词库词性标注类别

词语	词类
领导人姓名	nL
国家名	ct
前职务	pof
后称谓	pob
普通姓名	nr
区别词（如“副”）	b
地区名称	ns

假设待分词的文本为 $T=S_1S_2\ldots Sn$，则分词优化算法描述如下：

算法 4. 分词优化算法

Step1： 遍历文本 T，若 T 中出现 UserWord 中的词，将其替换为“‘ddcc’+序号+‘ ’”；

Step2： 调用分词程序对替换后文本进行常规分词处理；

Step3： 将分词后文本中的形如“‘ddcc’+序号+/x”的词，替换为 UserWord 中对应的词形如“‘词’+’/’+‘词类’”；

Step4： 输出结果。

注：“ddcc”是“单独成词”的拼音缩写，表示一个单独成词的字符串。

4.3 政治性差错侦测模型

真实的政治领域文本中的差错多发生在语义级，这类错误类型比较固定，但是具体的错误形式却五花八门。通过分析大量的政治新闻语料和相关的政府文件以及真实政治性差错语料，提取制定了相关的错误推理规则库，针对不同类型错误采取不同的规则进行分类侦测。

通过对《人民日报》语料和互联网时政文章的统计分析，提出政治性差错校对的一般形式化规则模型如下：

$$S(K,I,T,B)+DC_i(K,I,T)\rightarrow O(K,I,T,C) \qquad (2)$$

句子分词后存入字符串数组 StringArray，同时将关键信息存入初始集 S(K,I,T,B)，其中 K 是 n 元组 K=($K_0K_1K_2\cdots k_n$)(n 为大于 2 的整数)，K_i 是句中各类政治类关键词集合，K_0 是领导人姓名集合,K_1 是领导人职务集合，K_2 港澳台术语集合，若 K_i 未包含元素，则 K_i=Ø；I 为 K 中词语在数组 StringArray 的序号集合；T 为传递性动词，若句中不存在，则 T=null；B 为初始文本字体颜色，表示黑色。DC_i(K,I,T)为规则函数集，下标 i 对应初始集 S 中关键词集 K_i 的下标。O(K,I,T,C)为 S(K,I,T,B)在规则集 DC_i(K,I,T)的作用下输出的文本信息，其中 C 为输出文本中字串的颜色集，C=(黑色,黄色,红色)，文本中黑色字串表示字串不存在错误，黄色表示可能存在错误，红色表示存在错误。

规则集的具体形式如下：

$DC_0(K_0,I,T)_1$：T=null，K_0 包含于领导人顺序库，且 K_0 中元素个数大于 1，若 I 中元素的数值大小次序符合领导人顺序库中次序，C=黑色；否则，C=红色。

$DC_0(K_0,I,T)_2$：T=null 且 K_1 中元素 $K_{0,i}$ 不属于领导人顺序库，若 $I_{0,i}=\max(I_{0,1}I_{0,2}\cdots I_{0,m})$，且 I_1 中其他元素的数值大小次序符合领导人顺序库中次序，C=黑色；否则，C=红色。

$DC_0(K_0,I,T)_3$：若 T!=null， 取文本中 T 之后的内容按照规则 $DC_1(K_0,I,T)$ 和规则 $DC_2(K_0,I,T)$ 进行处理。

4.4 面向政治领域的文本校对方法的具体实现

本校对方法的实现采用两级侦测和分类侦测的方法。两级侦测分别为：第一级常见错误和引号词侦测，第二级政治性差错侦测；分类侦测则是按照政治性差错的类别侦测。分级分类侦测的文本校对流程图如图1所示。

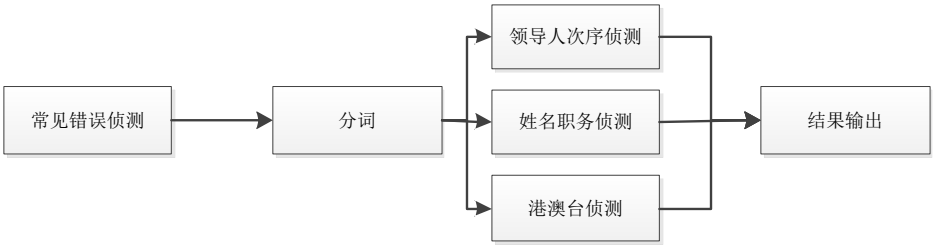

图1. 分级分类错误侦测的文本校对流程图

由于对字词级错误的侦测已有表深入的研究[7]，故对姓名职务对应错误中的第⑤种情况不作考虑，本文只针对语义级的错误，为了简单处理，假设待校对文本中每句话至多含有一处错误。

规则1：每一级和每一类错误都统一存放在错词组Error（words,num,index,type），记录其所在句子的序列号num，句中的位置index，和错误类型type，最终处理后的文本中错误字体颜色标注为红色。

规则2：在进行错误侦测时，若句子在上一级已侦测出错误，则终止本句侦测，跳转到下一句。

规则3：对于实际文本中单句存在多处错误的情况，采用重复侦测的方法，人机交互修

改标注的错误，直至无标红字体（即计算机认为文本中不存在错误）。

算法5是采用分级与分类侦测相结合分析具体例子“中共中央总书记、国家主席、中央军委主席习近平在李克强总理的陪同下来到中华世纪坛。”的描述。

算法5. 分级分类错误侦测算法

Step1. 输入待查错文本，遍历文本，将引号词库内包含的，但句中未加引号的词条记入错词组 Error（words,num,index,type）；

Step2. 分句处理文本，假设第 k 句是上述例句，检查 num 是否包含 k，若是，则处理第 k+1 句；否则，对句子进行分词预处理，分词后为：中共中央总书记/pof 、/wn 国家主席/pof 、/wn 中央军委主席/pof 习近平/nL 在/p 李克强/nL 总理/pob 的/ude1 陪同/vn 下/f 来到/v 中华/nz 世纪/n 坛/ng 。/wj

按空格切分放入字符串数组如下：

0	1	2	3
中共中央总书记/pof	、/wn	国家主席/pof	、/wn

依据词性标注信息提取关键词分类放入 S(K,I,T,B)，K_0(习近平，李克强)，K_1(中共中央总书记，国家主席，中央军委主席，总理)，I_0(5，7)，I(0,2,4,8)，T=null；

Step3. 按政治类关键词的类别分别应用对应的规则集 DC(K,I,T)进行侦测：

（1）若 K_0 至少包含两个元素，则使用领导人次序规则集 DC(K_0,I,T)，假设 K_0 内元素顺序差错，则将错误词条放入 Error[(习近平,k,5,T_0),(李克强,k,7,T_0)]，转到 Step4；否则转到（2）；

（2）若 K_0 和 K_1 不为空，判断 K_1 内职务与其修饰的 K_0 内姓名是否对应，若不对应则将职务和姓名错词放入 Error，转到 Step4，对应则看职务是否正确，错误则将职务职务加入 Error，转到 Step4； 否则跳转（3）；

（3）若 K_3 不为空，则应用港澳台规则集，因为该规则集包含规则较多不再详述，具体步骤类似（1）和（2）。

Step4. 若已扫描完所有文本，则转 Step5；否则处理第 k+1 句，跳转至 Step2；

Step5. 将错词组Error内记录的内容，换算为该错词在整个文章中的索引存入O(K,I,T,C)，标红显示输出。

5 实验结果分析

5.1 测试集的构建

由于报社期刊对于政治性校对的严格把关，真实文本中的政治性差错的相对较少，我们粗略统计了四种错误的分布比例：领导人姓名错误占20%，领导人顺序错误占10%，领导人姓名-职务对应错误占20%，涉及港台澳的政治性差错占40%。为了更好的模拟真实错误。我们选取1000个涉及政治性关键词的句子，句子各类型的分布比例符合我们的统计比例，首先进行分词处理，然后按照以下原则构建测试集，把1000个正确的句子和1000个错误的句子合在一起构成2000个句子的测试集ZZ。

（1）每个句子中只包含一处错误。

（2）除领导人姓名外，不对其他词进行替换单字、加字或删字的处理。

5.2 结果分析

利用ZZ测试集我们对错误侦测模型进行了测试，并做了语义搭配模型[9]的对比试验，具体试验结果如表3和表4所示。由表3中的比较可以看出，本实验在准确率、召回率以及F值方面的表现都较为突出。语义搭配模型主要检测词语间的语义搭配是否合理，虽是针对政治性领域文本进行训练，但整体的召回率和F值偏低，当然这与测试集的错误类型有一定关系。

表3. 实验结果

试验方法	准确率	召回率	F值
语义搭配模型	65.50%	34.50%	45.20%
本实验模型	80.50%	65.50%	72.23%

表4. 实验结果

	句子数	发现错误	实际错误	正确发现	召回率	准确率	误报率
正确的句子	1000	130	0	0			13%
政治性差错	1000	684	1000	684	68.4%		
总量	2000	814	1000	655	65.50%	80.50%	19.50%

通过表4可以看出，本文方法准确率较高但召回率偏低，造成这种结果的原因有：

（1）受限于错误语料的匮乏，规则库的规模偏小，一些错误模式并未在规则库中登录，所制定的规则没有考虑到语言中那些经验性的、小粒度的知识，覆盖不了各种复杂纷繁的语言现象。

（2）政治性新闻文本虽具有一定的用语规律，但其错误却非常分散，较难总结，且很多错误句子本身没有问题，只是语用方面的错误，这类错误的计算机自动校对还是很难实现的。

（3）尽管在分析涉及港台澳问题时首先使用文本分类方法，判定其是否属于涉港台澳的文章，但在具体文章的错误侦测时，仅考虑了词语的上下文，而未考虑句子的上下文，因而对“习金平将出席博鳌亚洲论坛年会”这样的句子，将无法判定“习金平”是否有错，因为他可能是一个企业家。后续的工作进一步收集新闻领域政治性差错方面的语料，补充完善规则库，对目前的方法进行改进。

参考文献

[1] 桂红星，陈 晖．报纸重大差错的成因及防堵[OL]，2006.8.29, http://www.cnhubei.com/200608/ca1147130.htm，

[2] 王燚．基于场景化知识表示的自然语言处理及其在自动文本校对中的应用[D]．成都：西南交通大学博士论文，2005.10

[3] 王亚东．消除报刊政治性差错需要注意的几个问题[J]．吉林省教育学院学报， 2012,Vol.28(2):125-126

[4] 郭爱民．书报刊中常见政治性差错例析[J]．科技与出版，2006（5）：50-52

[5] 新华社．新华社新闻报道中的禁用词（第一批），新闻阅评动态，第315期

[6] Li M, Zhang Y, et al. Exploring Distributional Similarity Based Models for Query Spelling Correction[C]. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL.2006:1025-1032.

[7] 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，2005，Vol.20(4):1-8.

[8] 李蓉.一个用于 OCR 输出的中文文本的拼写校对系统.中文信息学报，2009，Vol.23(5):92-97

[9] 管君,谢伟,张仰森.基于多知识源的语义搭配知识库的构建及应用[J].计算机工程与设计.2013，Vol.34(6):2136-2140.

作者简介：通讯作者：张仰森（1962），男，博士后，教授，主要研究领域为中文信息处理、人工智能，Email: zhangyangsen@163.com；

唐安杰（1990），男，在读研究生，主要研究领域为中文信息处理，Email:t_anjie@qq.com；

张泽伟（1988），男，硕士，主要研究领域为智能信息处理，Email:zhangzewei2008@163.com。

 张仰森  唐安杰  张泽伟