汉语文本自动查错与确认纠错系统的研究

慕勇 孙才 罗振声
清华大学中文系（100084）

摘要 本文介绍了一个针对大规模真实文本的计算机自动查错和确认纠错系统．它的基本思想是基于汉语言分析的多层次文本查错，然后以人机交互的方法进行纠错．内容涉及汉语自动分词、自动词性标注、句型成分分析、语法语义检查等问题．

关键词：文本自动查错与确认纠错 自动分词 自动词性标注 句型成分分析 语义检查

Research on automatic checking and confirmative correction of chinese text

Mu Yong Sun Cai Luo Zhengsheng
Dept. of chinese language and literature ,TsingHua University(100084)

Abstract In this paper, an automatic checking and confirmative correction system for large-scale real Chinese text is proposed. The basic idea of it is based on the multi-level analysis of Chinese language to find the errors in the text, then use the computer-aided method to correct them. The content refers to automatic segmentation,automatic tagging, analysis of sentence pattern's components and semantic checking.

Keywords: automatic checking and confirmative correction, automatic segmentation, automatic tagging, grammatival analysis of sentence pattern, semantic checking.

引 言

随着中文信息技术，特别是现代照排技术的发展，正在历史性地改变着出版业的面貌，电子出版物应运而生，专业录入行业迅速兴起．然而与之形成极大反差的是汉语文本校对工作仍然停留在原始的人工作业上，效率低，强度大，周期长．此外，语音识别和汉字识别也需要后处理．汉语文本的查错与纠错系统的研究已成为一项亟待解决的紧迫课题．

为此，近年来一些学者，特别是企业界开始关注这项工作，初步进行了一些实践．但由于汉语本身的特异性，至今尚未取得较大进展，距实际应用尚有一定距离．从长远看，信息化是社会发展的总趋势，其必将会出现越来越多的各类文本，如电子书、电子报纸、电子邮件、办公文件等，如何保证这些文本的正确性，将会显得越来越重要．

然而，汉语文本查错与纠错的研究，又是一项高难度课题．因为，研究的对象是建立在错误文本的基础上的．

一、汉语自动查错和纠错中的困难

汉语文本自动查错与纠错的困难主要在于:

1.汉语本身具有特异性，汉语的分析十分困难。

当前自然语言技术从理论到实践都还不足以对语言进行全面的正确分析，特别是面对大规模的真实文本时，这一点显得更为突出。正是由于汉语本身的这种特点，人们发现对它的分析较西语困难得多。这一点是众所周知的：即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。

2.面向大规模真实语料的错误文本，分析难度更大。汉语文本纠错正是针对真正意义上的大规模真实语料的错误文本。在面向正确文本分析技术尚未取得有效进展的今天，面向错误文本的分析与查错纠错，其困难是显而易见的。

为此，本文试图集当前自然语言技术一些较成熟的手段，并采用规则与语料库语言学相结合的方法，以及实施针对具体应用的策略，进行多层次多角度的分析查错。目前已初步取得一些进展。

二、汉语文本常见错误分析

1.根据错误的成因，汉语文本错误大致上可分为录入错与原稿错两大类。

1) 录入错误：指在文稿形成后，以各种方式录入计算机时所产生的错误。

①错别字：如击错键、重码或者联想输入中误选，可形成一些错别字.

【例1】原句：FoxPro都要计算该表达式的值。

录入为：FoxPro都要计算该表过式的值。

②漏字、多字与串行：如录入员由于疲劳跳过一个或几个字，或者多余的删字操作，造成缺字；而多余的击键动作，在联想方式下则可能造成多字错误。

【例2】原句：让FoxPro使用常规的方法来计算表达式。

录入为：让FoxPro使用常规则的方法来计算表达式。其中“则”字可能是多用了联想造成的。多字的另一种常见现象是同一字的重复。

【例3】括弧中是实际的表表文件名。

此外，也可能出现多余的标点符号。

缺字错误可能是漏一个，也可能连续漏多个字甚至串行，如果将标点符号漏掉，则可能造成超长句。

2)原稿错误：文稿在形成过程中由作者疏忽而形成的错误。除上述情况外，主要是一些语法错误。

①搭配不当：包括主谓、述宾、修饰语和中心语、关联词语及固定结构搭配不当等。

【例4】他的晚年，仍然是精神焕发，写下了不少好作品。

句中的主语应为“他”而不是“他的晚年”。

②位置不当：主要是词序不当，多出现在附加成分上。

【例5】这个厂不重视技术革新，产量不是比别人低，就是质量比别人差。

③结构残缺：包括句子主要成分如主、谓、宾的残缺以及残缺必要的虚词成分等。

【例6】通过这一事件，给了大家深刻的教训。

句中“这一事件”本来是主语，被加上“通过”后，句子就缺了主语。

④结构混乱：主要是不同句式的杂糅。

【例7】学习任务再重，越要注意锻炼身体。

句中的“再”和“越”分别是两种句式。

语法错误一般比较隐蔽，造成的问题多在句子结构上或语义上。

当然，录入过程版式错误也是常见的，这类错误较易于纠正。目前，本文主要针对文本中的文字性错误，所以，这些错误并非本文当前工作的重点。

2.根据语法、语义分析，常见文本错误有以下几种类型：

一个字或词之所以被认为错了，是因为它与其所在的上下文环境不相适应，即语法、语义关系搭配不当造成的。

1)构词错误：错字、缺字或多字破坏了原文词的结构，出现所谓“非词”现象，包括出现了一些不能单独做词的单字。根据我们统计，在GB2312-80的6763个国标汉字中，能够单独做词的只有大约2600个，其它的汉字一般不能单独成词。

2)句法错误：某些错误虽破坏了原词的结构，但却能与其前后字构成词，或者该单字本身就是词，从而并不违背构词法，但却可能破坏句子的整体结构，造成语法错误。

①词性搭配错误

根据我们对清华大学TH语料库的2级语料统计结果表明：汉语的词性邻接搭配关系较灵活，但也有一定的规律性，请看下例：

【例3】只有一第记录是当前记录。

句中的“第”本为“条”，“一条记录”的词性序列应为“mx qni ng”，其中，mx为系数词，qni为个体量词，ng为普通名词。这是一种常见的词性搭配，但由于上述错误，其词性序列变成了“mx maf ng”，其中，maf为前助数词，而它只能出现在数词之前。

此外，较为典型的例子是某些词类受一定限制不能出现在句首或者句尾。如语气词“吧”，“吗”，“呢”等不能出现在句首；而中置词如“和”，“并”，“或”，“分之”，则不可出现在句尾。这类错误可以通过检查词性邻接关系发现。

②关联词语搭配错误

前面已经提到，汉语中存在着介词结构一类的固定格式，其通常由两个必备的词形成某种搭配框架，主要格式有：主从连词结构类，如“因为…所以”，“虽然…但是”等；介词结构，如“在…中”，“从…以来”等；其它，如“越…越”等。

在查错时，如果前件与后件搭配不当，则可对其纠正。

③句型错误

所谓句型，是对具有相同的句法、语义和语用特征的句子的概括。通过对句子的句型与句型成分分析，可以发现可能发生的诸如句子成分残缺、句子成分位置不当等一类错误。

3)语义搭配错误

汉语的词与词之间、短语与短语之间还存在语义搭配关系问题。语义的研究是一个大的课题。本文只关心在句型成分之间，以及句型成分短语内部的语义搭配关系的分析与检查。如：动词的施事和受事搭配关系与限制；句型成分短语内部词与词之间搭配等.

如处理形如vgn+ng1+的+ng2一类歧义进行的检查。

三、 汉语文本自动查错与纠错的策略

1.机器自动查错与人工确认纠错相结合的基本策略

汉语文本的自动查错与确认纠错的研究是一项艰巨的任务。从当前自然语言处理技术发展的实际情况出发，我们采取的基本策略是先用计算机自动查出文本中的错误，然后通过人机交互方式纠错。由于文本中出现错误的句子一般远远少于正确的句子，所以，查错工作量远远大于纠错的工作量。因此，由计算机进行高速查错并提供可能的纠正方案，然后由人工进行确认纠错的方案在目前看来是可行的。

2.多层次查错、纠错的策略

从以上讨论看，文本中常见错误可能发生在词法、句法以及语义等多个层次上，因此单纯靠某一层次的方法不可能最大限度地解决问题。例如，即使自动分词检查完全正确，也不说明文本中没有错误。因此，本文从实际出发，采用了多层次的分析查错纠错的方法。其特点是:

1)多层次查错的思想与现代语言学的层次分析方法相吻合，并可更好地把自然语言技术现有成果应用到查错处理中。

2)层次化的方法实际是对问题的一种细化，类似于动态规化中将问题划分为若干子目标，并以较有效的方法解决子目标中的一些问题。

3.规则方法与语料库统计方法相结合的策略

规则方法与语料库统计方法相结合，是当前计算语言学发展的主要趋势。94年中期，我们已建成了“清华大学ＴＨ大型通用汉语语料库系统”，并通过了专家鉴定。该语料库规模大，功能完善，为课题研究工作提供了强有力的帮助。我们的汉语文本自动查错与确认纠错的研究，正是利用了这一优势。例如:

1)通过对约250万库存 2 级语料的统计，获得了自动标注用的两项统计模型参数：二元词性标记同现关系矩阵以及某词出现为某词性的概率。并以此对文本进行自动标注，取得了很好的效果。

2)在分词阶段，对人名的辩识，采用语料库统计与规则相结合的方法。在我们的语料库中，存有本校在校职工与学生的名单，通过对它的统计，获得了有关中国人人名用字

的数据，包括单字人名，双字人名首字，以及双字人名末字的初步规律；此外，我们还利用了规则对人名的上下文进行分析，并加以处理。

4.面向查错的“粗分析”方法

查错的目标是尽可能地召回文本中存在的错误，这与一般语言分析有某些不同：

1)出发点不同。句法分析首先假设待分析句子是正确的，它意味着一个句子必然对应一棵句法树与一个语义网络；而自动查错则认为待分析句子可能是错误的，分析过程正是查找错误的过程。从这一点上讲，自动查错与确认纠错的难度要大得多。

2)分析的粒度要求不同。句法分析要求为输入句子建立一棵以词为叶子结点的句法树，否则认为分析失败；而自动查错并不试图建立这种句法树，这意味着并不要求解决句子中所有语法问题。从这一点上说，自动查错又较一般句法分析灵活。

3)句法分析中的某些歧义结构在自动查错与确认纠错分析中可以不作为歧义结构处理。例如，句子“这是一幢新职工宿舍”中的“新职工宿舍”，在句法分析中有两种可能的分析结果，但在自动查错的句型成分分析中把“新职工宿舍”分析为宾语，并不去追究该短语内部的细微结构。而在以后的语义检查中也只需找出这一部分的中心词“宿舍”，而不必关心其具体的修饰关系。

5.具体应用的针对性策略

鉴于现阶段语言分析技术尚不可能解决文本中的所有问题，在自动查错中应针对出现频率较高的具体问题采取针对性的措施。如在分析过程中，我们发现很多错字是由于输入过程中的疏忽造成的，而且与当时使用的输入法又有很大关系。如五笔字型输入法造成的错字形似，而拼音法音似等。针对性的措施在本文系统中取得了较好的效果。

四、实验系统简介

1.实验系统的结构

本实验系统共分为五大模块：

1)预处理模块，包括：

①文本整理：完成各种文本格式的转换，形成系统内部的标准文本供分析处理。在此过程中还可以发现诸如超长句子、不符合叠字词规则的重复字以及不合法的相连标点符号等错误。

②新词发现：发现文本中包含的新词，并经整理后形成针对一批文本的临时词典。

2)主处理模块，为系统的核心，完成对文本的多层次分析和检查。包括分词、词法检查、词性标注、词性检查、句型成分分析、句法检查、语义分析和检查。当发现错误时记录错误发生的现场，通过报错机制报错，并以人机交互方式进行纠错处理。

3)知识库，包括规则、词典以及其它知识。如二元词性邻接矩阵、人名用字知识等等。

4)知识库维护模块，完成对知识库的维护，保持其一致性、有效性，并消除冗余。

5)学习模块，完成误报时的机器学习以及从语料库中获取知识。

6)输入输出模块，完成菜单，文本输入和输出的管理。

2.实验结果与分析

为了有效地对实验系统进行验证，我们从清华大学出版社激光照排中心的未经校对的录入稿《FoxPro数据库》中抽取了三章作为测试文本。同时还获取了毛校稿以便与系统分析结果比较。测试文本中共包含20,176个汉字，经毛校发现其中有错误239处，除去版式错误。西文符号错误以及几个不能辨识的中文错误以外，纯文字性错误有102处。对系统进行测试的结果，共发现错误203个，其中91个为真正的错误。设:

召回率=文本中的错误被发现的比例

精确率=被判为错误并真正为错误的比例

则系统的召回率和精确率分别为: 89%和40%

以下给出部分运行实例:

【输入】FoxPro都要计算该表过式的值。

【输出】FoxPro都要计算该<表过式 建议改为 表达式>的值。

【输入】从而也将这两个表关联焉了。

【输出】从而也将这两个表关联<焉 构词错误>了。

（实际为 起来）

【输入】只有一第记录是当前记录。

【输出】只有--<第 记录 词性邻接错误>是当前记录。

（实际为 一条记录）

【输入】此时整个表达式也可部分优化的。

【输出】<此时整个表达式也可部分优化的。 句型错误>

（实际为 此时整个表达式也是可部分优化的。）

我们认为，在汉语文本自动查错与确认纠错中，召回率比精确率更为重要。这是因为，一般汉语文本都是相当大的，虽然上述实验表明，系统的误报达到了55%，但如果与整个文本的汉字数相比，则尚不足6‰；与此相比，89%的错误能够被发现，还是令人满意的。

参考文献

[1] 罗振声，郑碧霞，《汉语句型自动分析算法与策略的研究》，《中文信息学报》94年第二期

[2] 罗振声，顾海波，《基于词类和短语结构的汉语句法、语义分析系统的研究》，第四届全国现代语言学研讨会，94年，北京

[3] 白栓虎，《汉语自动词性标注系统的研究与实现》，清华大学计算机系硕士论文，1992.

[4] 郑碧霞，《汉语句型自动分析和分布统计模型的研究与实现》，清华大学硕士学位论文，1993.