模式匹配和句型成分分析相结合的语法错误自动检查

龚小谨 罗振声 骆卫华

清华大学人文学院计算语言学研究室 北京 100084

E-mail: gxj99@mails.tsinghua.edu.cn

摘 要：本文将中文文本的语法错误分为搭配错误和与句型成分相关的错误两大类。分别采用模式匹配和基于句型成分分析的方法进行检查，这两种方法的结合，可以同时考虑局部和全局的语法限制信息，并且降低了语法检查的复杂度。通过对实验结果的分析和评测，证明本文所述的方法是可行的。

关键词：语法错误、模式匹配、句型成分分析

Combining Pattern-Matching and Sentence-Parsing Methods for Automatically Detecting Syntactic Errors

Xiaojin Gong, Zhensheng Luo, Weihua Luo

School of Humanities and Societies Science, Tsinghua University, Beijing 100084

E-mail: gxj99@mails.tsinghua.edu.cn

Abstract: The syntactic errors in Chinese texts can be divided into two categories: matching errors and sentence-component-relevant errors. This paper proposes a combination method of a pattern matching algorithm and the analysis based on sentence components. This method considers both local and global syntax information and can reduce the complexity in detecting syntactic errors. Experiments indicate that this approach is feasible and advanced.

Keywords: syntactic error; pattern matching; analysis of sentence components

1 引言

目前在中文文本自动校对这个课题中，对字词级错误的检查已有比较充分的研究（文献[1]、[2]、[3]、[4]），相比之下，语法和语义校对的研究尚不够深入。

从语法上来说，由于汉语本身的特点和汉语理论研究的局限性，使中文语法校对的难度要高于英文。语法错误检查的困难在于：①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、汉语词序的灵活性等汉语本身的特点使得汉语的语法分析存在很大的难度；②文本校对的处理对象是错误的文本，在面向正确文本的语法分析尚未有效解决的今天，面向错误文本的语法检查的困难可想而知。

目前，中文自动校对系统中语法错误检查运用的方法主要有基于统计的方法（文献[5]、[6]）、基于规则的方法（文献[7]）等，它们在提高整个校对系统召回率和精确率方面起到了一定的作用。但这些系统的语法校对方法也存在着一些不足：①统计方法中使用的词类邻接矩阵的 n 元模型，只能反映局部的语法限制，而不包括长距离的语法限制；②规则方法中句法的自底向上或自顶向下的分析方法不能发现特定词的搭配错误，同时耗费的系统开销也很大。

本文尝试着挖掘语料中语法错误的信息和特点，提出了一种模式匹配和句型成分分析相结合的语法错误检查的方法，兼顾了局部和全局的语法信息，并在不遗漏有用语法信息的前提下，简化句法分析算法。实验表明，校对的召回率和正确率都比较令人满意。

2 中文文本的语法错误分析

语法错误通常是由于漏字、多字或作者本身误用词语等引起的局部或全局语法错误。这里考虑的多字漏字是指由此导致一个词变成另一个有效词，从而引起语法结构错误的情况。从错误产生来源看，可能是原稿中本身存在的，也可能是因录入、OCR 和语音识别等产生的。

我们通过对收集的错误语料中存在的语法错误情况进行分析，将错误分为两大类：

1 搭配错误

1） 相邻的词类搭配错误；如：

例 1：她很兴致地发问。（副词"很"不能直接修饰名词"兴致"）

2） 关联词搭配错误；如：

例 2：只有迅速提高科学文化水平，我们就能适应快速发展的社会。（"就"应为"才"）

3） 长距离的特殊词搭配错误

例 3：三位老同志们下午来学校。（指人的普通名词，如果前面有表示确定数目的数量短语修饰，名词后面不能加"们"）

2 成分相关的错误

1） 成分重复多余、成分残缺等错误；如：

例 4：这根棍子不常。（原来是形容词词性的"长"字错写成了副词"常"，使该示例在成分分析时找不到谓语）

2） 成分内部的错误；

例 5：这种认为学习好，是很不对的。（在本句的主语中，"这种"只能修饰名词性短语，但"认为..."是谓词性短语）

系统用两遍扫描对以上不同类型错误进行检查：第一遍扫描，使用基于模式匹配的方法检查搭配错误；第二遍扫描，使用以谓语中心词驱动的句型成分分析方法检查成分相关错误。

3 基于模式匹配的语法错误检查

3.1 错误规则的模式定义

模式（M）是在错误语料中总结出词或词性搭配错误的规则，并加以形式化描述的结果。我们将一个固定模式用 BNF 描述如下：

M ::= LR + RR + END_COND；

LR ::= RR ::= R；

R ::= < DW，LF，RF，OP，ST，FT>；

DW ::= TS | TS + < WS > | TS + !< WS > | [TS]；

LF ::= RF ::= F ::= ϕ | F_ITEM + { F_ITEM }；

OP ::= ϕ | OP_CODE | OP_CODE + < L_OP R_OP>；

END_COND ::= F；

其中，LR、RR、R 分别为左规则、右规则和规则；END_COND 为终止条件；规则 R 是一个六元组：DW 为驱动词类，由词类(TS)和特定词集(<WS>)组合而成，在分析过程中，如果被分析词的词类与 DW 中表示的词类 TS 一致，并且被分析词属于特定词集<WS>时，唤醒该规则；LF 为左格式，是指被分析词的上文环境；RF 为右格式，是指被分析词的下文环境；OP 为操作，当上下文环境同时满足时，根据操作代码 OP_CODE 和左操作数 L_OP、右操作数 R_OP 的定义执行相应的操作，并标记被分析词的语法标注 ST 和功能标注 FT。

如上文示例 3 的规则：指人的普通名词，如果前面有表示确定数目的数量短语修饰，名词后面不能加"们"。用定义的形式表示为：

mx::qni:::: kn<们>:ng:::::[-]-]-；-。-：-！-？-@]:

其中左规则为"mx::qni::::"，左规则中的 DW 为系数词(mx)，LF 为 ϕ，RF 为个体量词(qni)，OP、ST、FT 为 ϕ；右规则为"kn<们>:ng:::::"，其中的 DW 为名词后缀（kn），特定词为"们"，LF 为名词（ng），RF、OP、ST、FT 为 ϕ；终止条件"-》-] -；-。-：-！-？-@"，其中的"-"为转义符号，当被分析的句子遇到"》"、"]"、"："、"。"等表示句子结束的标点符号时该模式匹配完毕。

3.2 模式匹配算法

在基于模式匹配的查错中，我们采用无回溯的确定性分析算法，同时构造了一个用于存放被激活模式的栈 mstack。具体的匹配算法如下：

1. 输入已经经过分词和词性标注的句子 S
2. 读入模式集{M}，并初始化 mstack
3. 如果 mstack 为空，则转 4；否则转 5
4. 从模式集读入一个 M，匹配 M 中的 LR，如果匹配，将该 M 入栈；否则读下一个 M 直到模式集为空
5. 取 mstack 的栈顶模式，匹配 M 中的 RR，如果匹配，则转 6；否则将该模式 pop 出栈
6. 匹配 M 中的 END_COND，如果匹配，则在句中标记错误，结束；否则将该模式 pop 出栈，结束

4 基于句型成分分析的语法错误检查

用模式匹配方法只能发现第一大类的匹配错误，对和句子成分相关的错误检查无能为力。所以在对句子进行基于模式匹配的语法错误检查后，需要用句型成分分析的方法再检查一遍。

本文的句法分析结合了实验室以前开发的汉语句型自动分析系统（文献[10]），采用以谓语中心词驱动的分治（divide and conquer）策略：

1. 先根据规则识别出一个句子的主语、谓语等成分块（chunk）；
2. 用自底向上的方法对成分块进行短语结构检查；
3. 分析检查成分块之间的关系

4.1 谓词的识别与检查

在谓语的识别和检查之前，先对句子进行预处理，将一些内聚性很高的词语子串捆绑为语片。语片在这里的定义是：具有一定的内聚性、能够结合在一起承担某种句法功能的词或词性标记的序列。如数量名结构"mx + qng (qni, qnk , qnv 等) + <ng>"、介词与方位词短语构成的框式结构"p + … + f"等（文中出现的词性标注的代码见表 1）。经过捆绑，可以简化谓语的识别，提高分析的正确性。

表1：词性标注代码

Vy：动词“是”	vgo：不带宾语的动词	Vgd：带双宾语的动词	qnk：种类量词	f：方位词
Vh：动词“有”	vgn：带体词性宾语的动词	Vgs：带小句宾语的动词	qnv：容器量词	p：介词
Mx：系数词	vgv：带动词性宾语的动词	va：助动词	ng：普通名词	
Qni：个体量词	vga：带形容词宾语的动词	vc：补语动词	kn：名词后缀	
Qng：名量词“个”	vgj：带兼语宾语的动词	vi：系动词	rn：体词性代词	

在判断一个词是否在句中做谓语中心词时，需要考虑该词的静态属性（记录词本身的特征）和动态属性（记录该词上下文的特征）。

词的静态属性是考虑该词的词性是否满足充当谓语的条件。在系统中，将动词分为 vgo、vgn、va 和 vc 等 14 个小类，其中有一些类的动词必不做谓语，一些类的动词可作谓语。如：词性为 va、vc 的词一般不作谓语中心词，"下来"、"出来"等趋向性动词一般不作谓语中心词。

词的动态属性，主要考虑该词是否在先前绑定的语片中、前后是否有虚词"的""得""地"等情况。词的动态属性用规则的形式表示，在识别中根据规则判断。

谓语的识别和检查算法如下：

```
INPUT：句子 S = W1…Wi…Wn
1．FOR i=1 TO n
        IF 词 Wi 不在语片中
            根据规则判断 Wi 是否可充当谓语，如果是，则加入准谓语链 PredLink；
2．FOR w = PredLink->fisrtword TO PredLink->lastword
        根据优先级 VerbPred  > AdjPred  > NounPred，当同时存在优先级高的和低
        的准谓语时，将优先级低的准谓语从 PredLink 中删除；
3．  词短语规则合并 PredLink 中可以合并的谓语；
4．  IF PredLink->num = 1
        标记该 S 的谓语成分正确；
    ELSE IF PredLink->num =0
        标记该 S 为谓语缺失
```

4.2 句子其他成分的识别和检查

从句子的结构来看，主语和状语都在谓语前面，宾语和补语在谓语后面。通常，状语和补语在句子中有比较明显的边界标志，结构也比较简单。

因此，这里采用句型分析系统中的策略：先识别状语成分，把谓语前面除状语以外的其它部分作为主语处理。在确定状语后，谓语前面如果已没有成分剩下，判断它是否满足无主句或省略句的要求，如果不是，则标记为主语缺失。

同样，在识别句子的宾语和补语时，先对谓语后面部分确定补语，剩余为宾语。在确定补语后，谓语后面若已没有成分剩下，判断谓语是否具有必须带宾语特性，若是，则标记为宾语缺失。

4.3 成分块的检查

以上的成分识别和检查是从句型结构的角度判断是否存在错误，对于每个确定了边界的成分块，尚需要进行成分内短语结构的分析检查。这里，主要对主语和宾语成分进行检查。我们将短语划分为名词性短语（NP）、动词性短语（VP）、介词性短语（PP）和单句型短语（DJ）等 10 类。使用上下文无关文法描述的短语规则对成分块进行检查，检查分为两步：

1）使用规则对成分块进行自底向上分析，看能否形成一个短语分析子树，通过分析子树的产生，检查短语结构是否正确；如上文示例 5：*这种认为学习好，是很不对的*。

经过成分确定后，把"这种认为学习好" 确定为主语，在对主语进行结构分析的结果为：

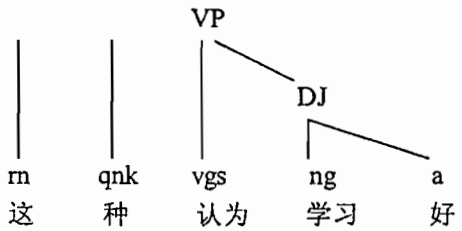

图1：语法分析结果

从该分析结果看，通过规则无法将成分块"这种认为学习好"归结为一个短语结构，从语法上看，"这种"只能修饰名词或名词性短语，而"认为学习好"是动词性短语。所以对不能生成顶结点（特指短语子树的根结点）的成分块进行错误标记。

2）检查短语分析子树的顶结点是否满足充当该成分块的要求，如：充当主语和宾语的通常是名词性短语等。不满足要求的，对该成分块进行错误标记。

5 实验结果及其分析

本文的自动语法检查系统是以经过分词和词性标注的中文文本为输入的，同时包含了4个规则库共522条规则，见表2：

表2：规则库说明

规则库		规则条数
错误模式规则		56
语片捆绑规则		153
成分识别规则	谓语识别规则	115
	状语识别规则	42
	补语识别规则	27
短语分析规则		129

系统使用的三个评价参数为：

召回率= 系统正确发现的错误数/ 文本中的错误总数

正确率= 系统正确发现的错误数/ 系统发现的错误总数

误报率= 系统发现的虚假错误数/ 系统发现的错误总数

在评测中共抽取了249个句子，其中有的存在语法错误，有的是正确的。测试结果如表3：

表3：测试结果

	句子数目	发现错误	正确发现	实际错误	召回率	误报率	正确率
正确的句子	120	15	0	0		12.5%	
搭配错误的句子	108	99	89	108	82.4%	10.1%	89.9%
成分错误的句子	21	19	13	21	61.9%	31.6%	68.4%
总量	249	133	102	129	79.1%	23.3%	76.7%

从测试结果看，无论是搭配错误还是成分错误，都存在一定的漏查和误查。其原因主要有：

1） 搭配错误中的错误模式覆盖不全，一些错误模式未在模式库中登录，导致搭配错误漏查；

2） 在模式库中登录的每一条错误模式，在检查错误搭配的同时，也可能会对正确的文本产生影响，使一些正确的搭配也被误查为错误。这是基于规则的方法本身的特点，需要通过实验对规则的信息粒度不断地进行调整，才能最大限度的减少误查；

3） 句型成分分析本身的正确率不是很高，这些错误在句型结构检查的时候可能不会产生很

大的影响，但谓语识别的不正确，成分边界确定的错误会直接导致成分块检查的错误；

4） 短语分析中规则的不完全导致成分块检查错误。

6 总结与展望

本文描述了一个中文文本自动语法检查系统，该系统有几个主要特点：

1） 采用了模式匹配和句型成分分析相结合的方法。这种方法同时考虑了语法的短距离和长距离的限制信息，提高了系统的语法检查能力。

2） 在句型成分分析中，采用 divide and conquer 算法的思想，将一个完整的句法分析分解为成分识别、成分块内结构分析等几个相对简单的子问题，有效降低了问题的复杂度，便于程序运行效率的提高。

3）在模式匹配中，手工收集和编写错误模式的规则。通过对正确和非正确的语料进行测试，调整每条规则的约束范围，尽可能地减少由此引起的误查。

中文文本的自动语法检查是一个比较难的课题，以后还可从以下几方面作尝试和努力：

1） 提高句型成分分析的正确率。这样可以有效地降低语法检查的误判率。

2） 继续挖掘语料中语法错误的信息和特点，扩大错误模式库。

3） 错误模式的自动获取。随着错误语料的增加，手工收集和编写错误规则的工作量会越来越大，将机器学习功能增加进来，可以更好的发现错误规则。

参 考 文 献

[1] 黎勇，孙才，罗振声，汉语文本自动查错与确认纠错系统的研究，计算语言学进展与应用，清华大学出版社，1995.10

[2] 罗振声，孙才，汉语文本校对字词级查错处理的研究语言工程，清华大学出版社，1997.8

[3] 张照煌，中文错别字自动订正方法初探，Communications of COLIPS. VOL.4. NO 2, DEC 1994, 143-149

[4] 张仰森，丁冰青，基于二元接续关系检查的字词集自动查错方法，中文信息学报，2000.8

[5] 于勐，姚天顺，一种混合的中文文本校对方法，中文信息学报，1997.5

[6] 张磊，周明，黄昌宁，基于 Winnow 的中文自动校对方法

[7] 王芸，何克抗，基于语法规则的计算机汉语文稿校对系统的设计，北京师范大学现代教育技术研究所

[8] KAREN KUKICH, Techniques for Automatically Correcting Words in Text. ACM Computing Surveys. Vol.24, No.4. December 1992

[9] Atwell E., Elliott S., Dealing with Ill-Formed English Text. In The Computational Analysis of English: a Corpus-Based Approach, De.Longman. 1987

[10] 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，1994.2