计算机汉语文稿校对系统

易蓉湘　何克抗
(北京师范大学现代教育技术研究所　北京　100875)

摘　要　本文对汉语文稿中常出现的错误进行了分析, 给出了错误规律的形式化表示, 并阐述了错误的识别方法和校正方法, 最后讨论了汉语文稿校对系统的初步实现

关 键 词　校对, 中文信息处理, 句法分析

中图法分类号　H1; TP391

PROOFREADING CHINESE MANUSCRIPT WITH COMPUTER

YI Rong-Xiang　HE Ke-Kang
(*Institute of Modern Education Technique, Beijing Normal University, Beijing* 100875)

Abstract　This paper analyzes errors that often appear in Chinese manuscript, and gives a method to recognize and correct the errors in several layers such as morphology, syntax and semantics. The implementation of a computer-based proofreading system for Chinese manuscript is also discussed.

Key words　proofreading, Chinese information processing, syntactic analysis

Class number　H1; TP391

0 引　　言

校对, 是指按原稿核对抄件或副样张, 看有没有错误. 人工校对时, 由人首先阅读了解原稿, 在此基础上判断并标记录入文稿中与原稿不同的错误, 以便后期修改. 计算机校对则不同: 首先, 计算机没有原稿作对照; 其次, 计算机处理的对象是电子文件; 另外, 计算机校对的错误与人工校对的含义不同, 人工校对的错误是指与原稿不同之处, 而计算机校对

收稿日期: 1995-01-18; 修回日期: 1996-05-04. 本课题得到国家"863"高技术发展项目的资助. 易蓉湘, 1969 年生, 1994 年于北京师范大学无线电电子学系获硕士学位, 现主要从事自然语言理解领域的研究. 何克抗, 1937 年生, 教授, 博士生导师, 本刊编委, 1963 年于北京师范大学无线电电子学专业研究生毕业, 主要研究领域: 中文信息处理, 自然语言理解和智能化辅助教学.

的错误则指不符合现代汉语的语法和语义之处

1　汉语文稿校对原理

1.1　错误的分类

人工校对过程中要校对的错误主要有两类: 一类是文字错, 一类是排版错. 文字错主要是一校中要消灭的多字、漏字、错字等差错, 二校中要校正的重点: 象形和错别字; 排版错则是指除文字错以外的格式、版面、图表、序号、注解等方面的错误. 在我们的汉语文稿校对中, 主要是针对前一类——文字错, 所以必须假设要校对的文稿中没有排版错误.

在这里我们用两条标准对录入文稿中的文字错误进行分类: 一条标准是根据汉字的增、改、少等的出错现象来分类, 另一条标准是根据错误在自然语言处理过程的不同阶段所导致的结果来分类.

根据汉字的增、改、少现象, 我们将录入文稿中出现的文字错误分为: ①CH 错误: 包括字、词、句的 CH 错误, 是指一个字、词或句变成了另外的字、词或句; ②DE 错误: 少了一个字、词或句; ③N 错误: 多了一个字、词或句, 主要表现为多字、词和句的重复录入.

根据错误在自然语言处理过程的不同阶段所导致的结果, 我们还可将录入文稿中出现的错误分为: ①词法错误: 词法分析可发现和处理不符合汉语词法规律的错误; ②句法错误: 句法分析可发现和处理不符合汉语句法规律的错误; ③语义错误: 语义分析可发现和处理不符合语义规律的错误; ④语用错误: 语用分析可发现和处理不符合上下文及语言交际环境(即语用规律)的错误. 参看文献[1]至[3].

CH, DE 和 N 错误是文稿中所产生错误的直接表现方式, 而词法、句法、语义和语用错误则是文稿中出现错误所导致的结果. CH, DE 和 N 错误均可导致词法、句法、语义和语用错误.

1.2　错误规律的表示

通过对大量错误现象的观察, 我们发现, 错误产生的最小单位是字. 校对时, 词法分析所处理的最小单位是字, 所以在词法分析阶段, 错误规律表示为字到字的映射. 如: 戊→戌(CH 错误); 戊→λ(N 错误); λ→戌(DE 错误).

句法和语义分析所处理的最小单位是词, 所以, 在这几个阶段, 错误规律表示为词到词的映射. 为此, 需要把总结出来的基于字的错误规律转化为词的规律. 同时, 由于这几个阶段的分析是采用基于产生式的分析方法, 因此, 还需要在产生式中增加词的信息, 把词的错误规律和产生式规则中的词相结合, 使扩充了的产生式能够反映出汉语文稿中的错误规律. 比如, 我们有一条句法语义分析用的产生式规则:

S→= PRN + 被+ NN + VC+ VH→

它所描述的是一种"被"字句, 如下句: 他被小贩骗过一回.

又, 存在基于"被"字的如下错误规律:

被→披(CH 错误); λ→被(DE 错误); 被→λ(N 错误).

在此,"被"字是词, 所以错误规律不需要进行字到词的转换. 那么, 在句法语义校对时, 错误规律的表示将是错误的基于词的规律在分析用规则中的落实:

S–= PRN + 披+ NN + VC+ VH –→PRN + 被+ NN + VC+ VH –(CH 错误);

S–= PRN + ^ NN + VC+ VH –→PRN + 被+ NN + VC+ VH –(DE 错误);

S–= PRN + 被+ 被+ NN + VC+ VH –→PRN + 被+ NN + VC+ VH –(N 错误).

这样, 这三条错误规则就可校正如下的句法语义错误:

他披小贩编过一回(CH 错误); 他^ 小贩骗过一回(DE 错误); 他被被小贩骗过一回(N 错误).

1.3 错误的校正

错误的校正, 和大多数其他的自然语言处理系统一样, 可以对输入汉字串分成词法、句法和语义几个层次进行处理. 但是与其他系统不同的是, 错误校正在处理的不同阶段其最根本的机制是相同的, 即根据错误发生的共同规律和表现形式来进行含错误汉字串的校正, 这是和其它系统的最大差别. 下面我们来看看有错误串的校正方法.

错误串的校正过程是把一个被确定包含错误的输入串变换成无错误的输出串. 即: 有错误的输入汉字串$L1(G')$, 通过错误校正系统, 变换成无错误的输出汉字串$L2(G)$. 这种串到串的转换影射, 其语言理论模型可以是句法引导的翻译系统. 形式上, 从语言$L1$到语言$L2$的翻译是一个在$L1\times L2$中的关系T. 对于在T中的翻译对(x, y), 我们可以说, 输入串x被翻译成输出串y. 值得注意的是, 一个给定的输入可能会有任意数量的互不相同的翻译. 假设输出语言$L2$由文法$G=(N, \Sigma, P, S)$描述和生成, 我们对终结符$a\in\Sigma$定义如下三个算子, 分别表示CH, DE 和 N 错误的结果:

$$\mathrm{CH}(a)=\Sigma-\{a\}.$$

$$\mathrm{DE}(a)=\lambda \qquad a\in\Sigma$$

$$\mathrm{N}(a)=\Sigma a\mid a\Sigma=(a1\mid a2\mid\cdots\mid al)a\mid a(a1\mid a2\mid\cdots\mid al)=a1a\mid a2a\mid\cdots\mid ala\mid aa1\mid aa2\mid\cdots aal$$

推广到Σ中的串x, 有:

$$\mathrm{CH}(x)=\begin{cases}\Phi(\mid x\mid<n).\\ \{y\mid y \text{ 在 } \Sigma^{*} \text{ 中}, x \text{ 中有 } n \text{ 个不同的位置出现 CH 错}, x \text{ 变成 } y\}(\text{其他情况}).\end{cases}$$

$$\mathrm{DE}(x)=\begin{cases}\Phi(\mid x\mid<n).\\ \{y\mid y \text{ 在 } \Sigma^{*} \text{ 中}, y \text{ 是由于丢掉了 } x \text{ 中 } n \text{ 个终结符得到的串}\}(\text{其他情况}).\end{cases}$$

$$\mathrm{N}(x)=\{y\mid y \text{ 在 } \Sigma^{*} \text{ 中}, x \text{ 中出现了 } n \text{ 个 N 错误而变成串 } y\}.$$

可见, 输入语言$L1$为:

$$L1(G')=\{y\mid y \text{ 在 } \mathrm{CH}^{n}(x), \mathrm{DE}^{n}(x) \text{ 或 } \mathrm{N}^{n}(x) \text{ 中}, x \text{ 在 } L2(G) \text{ 中}, n\geq 0\}.$$

对应于G的P中产生式$A\rightarrow a\alpha$(用格雷巴赫范式给出), G中产生式是(n简化为1):

无错误: $A\rightarrow a\alpha$; CH 错误: $A\rightarrow \bar{a}\alpha$(对$\Sigma-\{a\}$中的每个$\bar{a}$); DE 错误: $A\rightarrow\alpha$;

N 错误: $A\rightarrow \bar{a}a\alpha$, $A\rightarrow a\bar{a}\alpha$(对$\Sigma$中的每个$\bar{a}$).

所以校正$L1$中的CH, DE 和 N 错误, 可以同样构造一句法引导的翻译系统

$$JE=(N, \Sigma, \Sigma, R, S).$$

R中规则确定方法是: 对应于文法G中的每个产生式$A\rightarrow a\alpha$, 有正面的规则:

无错误: $A\rightarrow a\alpha, a\alpha$; CH 错误: $A\rightarrow \bar{a}\alpha, a\alpha$(对$\Sigma-\{a\}$中的每个$\bar{a}$); DE 错误: $A\rightarrow\alpha, a\alpha$;

N 错误: $A\rightarrow \bar{a}a\alpha, a\alpha$, $A\rightarrow a\bar{a}\alpha, a\alpha$(对$\Sigma$中的每个$\bar{a}$).

2 汉语文稿校对系统的实现

汉语文稿校对系统的处理分为词法和句法语义等层次进行校对, 见图1.

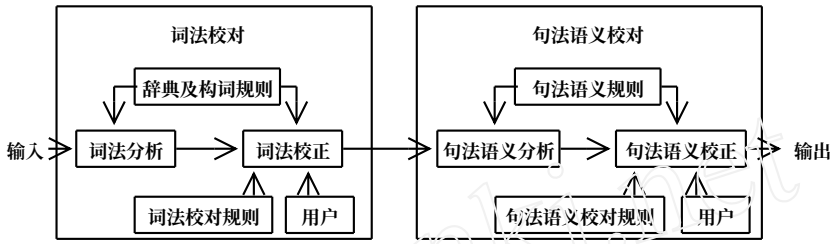

图1　汉语文稿校对系统结构框图

2.1　词法错误的校正

词法错误的校正可用前面所介绍的错误校正原理解决. 符合汉语词法规律的词可以被这样的文法 G 描述: $G = (N, \Sigma, P, W)$.

其中, Σ 为汉字集, 产生式集 P 由辞典中所有的词以及构词规则所表示和描述. 对 Σ 中所有汉字作 CH, DE 和 IN 运算. 这里, 我们将忽略极小概率事件, 对 CH, DE 和 IN 三种错误算子加以约束修改, 使之得到简化.

$\mathrm{CH}(w) = \{v \mid v \in \Sigma,\ v$ 和 w 易混$\}$　　$w \in \Sigma$ 中易混字;

$\mathrm{DE}(w) = \lambda$　　$w \in \Sigma$ 中易漏字;

$\mathrm{IN}(w) = \{u1w, wvj \mid u1 \in \Sigma,\ vj \in \Sigma, u1$ 易加在 w 前, vj 易加在 w 后$\}$

$w \in \Sigma$ 中前易加字 $u1$ 和后易加字 vj 的字.

在此基础上, 我们构造一个翻译系统 JE, 得到一个对词法层次上 CH, DE 和 IN 错误的不确定的校正. 我们知道, 录入员易出现的错误是有规律的, 原稿作者易犯的错误也是有规律的, 例如, "母"和"毋","戍"、"戌"、"戊"和"戒"等易混, 发生 CH 错误. 我们把这些规律进行搜集、归纳和总结, 就可使 CH, DE 和 IN 算子得到简化, 使我们的词法校对变得可行而有效. 由于词法规则存在歧义, 校正结果可能不是唯一的, 这时需要人工干预才可消除校正的不确定性. 在词法校正过程中, 除了辞典和构词规则外, 我们还需要一个能反映词法错误规律的库. 该库的来源有二, 一是从事手工校对的同志归纳总结出的常见校对错误中的词法错误部分, 在一般的校对手册中都包含有这部分内容; 二是通过上机录入实践以及分析汉字编码规律, 可以总结出录入员易犯的错误. 这类规律和汉字编码关系较大, 不同的编码方法, 错误规律是不同的.

2.2　句法语义错误的校正

为发现句法语义错误, 我们采用以谓语为中心、特征词优先的算法来进行句法语义分析, 将语义分析结合在句法分析中, 并用于引导句法分析.

算法所基于的规则采用形式化的产生式表.

产生式规则的形式为:

〈产生式规则〉: = 〈句法结构〉[(〈特征信息〉)]= 〈句法项〉{+ 〈句法项〉}

[ ]内的内容可有可无, {}内的部分可以重复出现 0 次以上.

句法结构表示为:

〈句法结构〉: = 〈结构头〉—[〈结构尾〉]

结构头表示该结构的句法功能, 结构尾标明该句法结构的特点, 或是有别于其他同结构头的句法结构的标号.

特征信息的形式为:

〈特征信息〉: = 〈特征名〉: 〈特征值〉

不同的句法结构的特征名和特征值都不同, 上层结构可以包含下层结构的特征信息 句法项分两种情况, 一种是特征句法项, 另一种是结构句法项

〈特征句法项〉: = 〈特征句法分项〉{ |〈特征句法分项〉|}[|]

〈结构句法项〉: = 〈结构句法分项〉{ |〈结构句法分项〉|}[|]

特征句法分项的形式为:

〈特征句法分项〉: = 〈特征属性〉[〈条件操作〉]

特征句法分项是具体词或词性, 特征属性即具体的词或某个词性, 其后可以有附加操作及测试条件:

结构句法分项的形式为:

〈结构句法分项〉: = 〈句法结构〉[〈条件操作〉]

句法项中的条件操作可以是: ①v: n 对音节的检查; ②o: , e: 记录并检查前后终结符是否相同; ③! "非"检查, 检查是否不具有某种属性; ④b: 动宾搭配检查; ⑤w: 谓语资格检查; … …:

句法语义错误的校正分两种情况, 一是错误语句对应的句法语义规则中出现的字词发生错误, 它们的校正可以通过构造句法引导翻译系统来实现; 二是错误语句对应的句法语义规则中未出现的字词发生错误, 它们的校正需要采用修改后重新对其进行句法语义分析的办法 对第一种错误, 我们的具体作法是: 构造句法引导的翻译系统:

$$JE = (N, \ , \ , R, S)$$

翻译系统的输入语言是被判断为出错了的句子集 含错误语言的产生式规则可以由汉语的句法语义规则经变换后得到, 具体作法是, 对每条规则中的具体字词充当的特征词进行 CH, DE 和 IN 运算, 将运算结果替换原规则中相应的特征词, 这样就得到了相应的输入语言的规则, 从而构成错误语言的产生式规则集 这样的翻译系统可以校正的句法语义错误只能是, 出现在该句子相应的句法语义规则中是具体词的特征词上, 亦即上文所说的第一种错误 由于在句法语义规则中, 绝大部分特征词的出现形式是词性这种终结符, 而我们知道, CH, DE 和 IN 错误本质上是以字为单位发生的, 所以, 构造句法引导翻译系统来校正这种发生在词性终结符上的错误是无能为力的 在这种情况下, 我们采取的办法是: 用易错字的 CH, DE 和 IN 运算结果修改原句, 再对修改后的句子进行句法语义分析, 成功了的替换句子就是该错误原句的校正结果之一: 从中选择一种发生概率最大的即可组合作为校正结果

参 考 文 献

1 刘开瑛, 郭炳炎 自然语言处理 北京: 科学出版社, 1991

2 孙茂松, 黄昌宁: 汉语中的兼类词、同形词类组及其处理策略 中文信息学报, 1989, 3(4)

3 陈志忠, 陈肇雄, 高庆狮 通用的自然语言词法分析机制 计算机学报, 1991, 14(2)