(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109460552 B
(45)授权公告日 2023.04.18

(21)申请号 201811268613.8

(22)申请日 2018.10.29

(65)同一申请的已公布的文献号
申请公布号 CN 109460552 A

(43)申请公布日 2019.03.12

(73)专利权人 朱丽莉
地址 646100 四川省泸州市泸县福集镇茂盛村一组217号

(72)发明人 朱丽莉 谭代龙

(74)专利代理机构 成都九鼎天元知识产权代理有限公司 51214
专利代理师 钱成岑

(51)Int.Cl.
*G06F 40/211*(2020.01)
*G06F 40/253*(2020.01)
*G06F 40/289*(2020.01)
*G06F 40/30*(2020.01)

(56)对比文件
CN 102541837 A,2012.07.04
JP H0981568 A,1997.03.28

审查员 宋晶晶

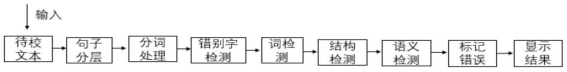
权利要求书6页 说明书16页 附图4页

(54)发明名称

基于规则和语料库的汉语语病自动检测方法及设备

(57)摘要

本发明公开了一种基于规则和语料库的汉语语病自动检测方法及设备，该方法包括：文本获取、句子分层、自动分词和语病检测；所述自动分词包括以下步骤：切分字符串步骤和自动分词步骤；所述语病检测为根据所述自动分词的结果和预先构建的语料库进行语病检测。语病检测包括错别字检测、用词不当检测、句法结构检测、语义表达检测，语病检测可以包括这四种检测中的一种或几种。语病检测包括的几种检测可以并列执行，也可以依次执行，且这几种检测的前后顺序可以根据情况进行选择。本发明从词法、句法、语义等角度进行设计，自动检测文本中的各类语病问题。




CN 109460552 B

1.一种基于规则和语料库的汉语语病自动检测方法,其特征在于,包括文本获取、句子分层、自动分词和语病检测;

所述文本获取为获取待校文本数据;

所述句子分层包括读取文本,获取文本中的句子数量,并将获取的文本划分为单句;

所述自动分词包括以下步骤:

正向切分字符串步骤,以单句为单位,获取单句字符串长度,从左往右依次切分出长度不等的字符串,即从第一个字符开始,依次切分出N个字符、N-1个字符、N-2个字符、…、两个字符、一个字符的字符串,N为单句字符串长度;

逆向切分字符串步骤,以单句为单位,获取单句字符串长度,从右往左依次切分出长度不等的字符串,即从最后一个字符开始,依次切分出一个字符、两个字符、三个字符、…、N个字符的字符串;

自动分词步骤,将切分出的字符串依次与预先构建的语料库中的固定式语料库和词语语料库比对,若匹配成功,则输出该字符串并标记序列号为1,2,3,...,z,若匹配失败,则将未被标记序列号的其他单个字符逐一与预先构建的语料库中的单字词语料库比对,若匹配成功,则为单字词,输出该单字词并标记对应的序列号为z+1,z+2,z+3,...,若匹配失败,则与预先构建的语料库中的非单字词语料库比对,若匹配成功,则为非单字词,输出该非单字词,若匹配失败,则保留该字符;

所述语病检测为根据所述自动分词的结果和预先构建的语料库进行语病检测;

所述预先构建的语料库包括固定式语料库、字母语料库、标点语料库、拼音语料库、繁简字语料库、语句语料库、词语语料库和单字词语料库;根据所述预先构建的语料库,进行的语病检测包括错别字检测、用词不当检测、句法结构检测和语义表达检测。

2.根据权利要求1所述的一种基于规则和语料库的汉语语病自动检测方法,其特征在于,语病检测包括错别字检测,所述错别字检测包括以下步骤:

错别字正向检测步骤,所述错别字正向检测步骤包括以下子步骤:

字母检测子步骤,判断切分字符串步骤中,切分出的字符串是否有数字或字母,若是,则将切分出的字符串与预先构建的语料库中的字母语料库比对,如果形式正确,则输出该字符串,如果形式错误,则输出该字符串并标记为(*);

标点检测子步骤,判断切分字符串步骤中,切分出的字符串中是否含有标点符号或特殊符号,若是,则将切分出的字符串与预先构建的语料库中的标点语料库比对,如果形式正确,则输出该字符串,如果形式错误,则输出该字符串并标记为(*);

拼音检测子步骤,判断切分字符串步骤中,切分出的字符串是否有拼音,若是,则将切分出的字符串与预先构建的语料库中的拼音语料库比对,如果形式正确,则输出该字符串,如果形式错误,输出该字符串并标记为(*);

繁体字检测子步骤,将待校文本与预先构建的语料库中的繁简字语料库比对,判断切分字符串步骤中,切分出的字符串是否有繁体字,若是,则获取繁体字数量,并将繁体字逐一提取,判断它是否属于引用或特别使用情况,若不是引用或特别使用,输出该繁体字并标记为(*);

单字词检测子步骤,将自动分词步骤中,判断为单字词的单字与下一单字组合,与预先构建的语料库中的语句语料库比对,若匹配成功,则输出该单字;将自动分词步骤中,判断

为非单字词的单字与下一单字组合，与预先构建的语料库中的语句语料库比对，若匹配成功，则输出该单字，若匹配失败，则输出该单字并标记为(*)；

错别字逆向检测步骤，以单句为单位，从右至左，将自动分词步骤中，判断为单字词的单字与下一单字组合，与预先构建的语料库中的语句语料库比对，若匹配成功，则输出该单字；将自动分词步骤中，判断为非单字词的单字与下一单字组合，与预先构建的语料库中的语句语料库比对，若匹配成功，则输出该单字，若匹配失败，则输出该单字并标记为(*)。

3.根据权利要求1所述的一种基于规则和语料库的汉语语病自动检测方法，其特征在于，语病检测包括用词不当检测，所述用词不当检测包括以下步骤：

用词不当分词结果获取步骤，获取待校文本句子分层结果及自动分词步骤的待校文本分词结果，所述分词结果为自动分词步骤依次输出的固定式、词语和单字词，并按固定式、词语和单字词在单句中排列顺序依次标记位置为1,2,3,...,Z；

用词不当检测步骤，所述用词不当检测步骤包括以下子步骤：

相邻词位置检测子步骤，结合预先构建的语料库中的词语语料库和单字词语料库中位置字段的标记，将待校文本中的词从自由与粘着两个角度自动标注位置信息；判断待校文本是否有定位词的标记，将待校文本中的定位词和相邻词与词语语料库和单字词语料库位置搭配规则对比，判断是否正确，如果错误，则输出错误词并标记为(*)；

相邻词词性检测子步骤，结合预先构建的语料库中的词语语料库和单字词语料库的词性字段的标记，将待校文本中的词自动标注词性信息；将待校文本中相邻词的词性与词语语料库和单字词语料库中的词性搭配规则进行匹配，判断待校文本中相邻词之间的词性是否能搭配，如果不能搭配，则输出错误词并标记为(*)；

相邻词语义搭配检测子步骤，结合预先构建的语料库中的词语语料库和单字词语料库的语义字段的标记，将待校文本中的词自动标注语义信息；将待校文本中相邻词的语义与词语语料库和单字词语料库中的语义搭配规则进行匹配，判断待校文本中相邻词之间的语义是否能搭配，如果不能搭配，则输出错误词并标记为(*)；

不相邻词语义搭配检测子步骤，采用互信息算法，对i和j这两个词，通过对比语句语料库，计算它们的互信息值$Q(i,j)=log_2\frac{P(i,j)}{P(i)P(j)}$，用Q(i,j)的大小判断这两个词语义的组合情况，当Q(i,j)>0时，表示i和j这两个词语义组合正确，输出i和j；当Q(i,j)＝0时，表示i和j语义组合不明确，则输出i和j并标注为(?)；当Q(i,j)<0时，表示i和j语义组合不正确，则输出i和j并标注为(*)；其中，i和j分别为待校文本单句中任意两个词，P(i,j)为i和j这两个词共现的频率，P(i)和P(j)分别为i和j这两个词出现的频率。

4.根据权利要求1所述的一种基于规则和语料库的汉语语病自动检测方法，其特征在于，语病检测包括句法结构检测，所述句法结构检测包括以下步骤：

句法结构分词结果获取步骤，获取待校文本句子分层结果及自动分词步骤的待校文本分词结果，所述分词结果为自动分词步骤依次输出的固定式、词语和单字词，并按固定式、词语和单字词在单句中排列顺序依次标记位置为1,2,3,...,Z；

标记虚词步骤，标记待校文本单句中虚词，与单字词语料库位置字段比对，自动标注虚词位置；

句法成分提取步骤，在分词基础上，从左往右依次切分出长度不等的字符串，即从第一

个词开始，依次切分出一个词、两个词、三个词、…、Z个词的字符串；计算切分出的字符串x为句子成分的概率$\tilde{P}(t|x)=\frac{freq(x,t)}{\sum_{x,t}freq(x,t)}$，如果$\tilde{P}(t|x)\geq 0.2$，则切分出字符串x，并标记成分名称；如果$\tilde{P}(t|x)<0.2$，表示该字符串x不是句中句子成分，则判断下一个字符串，直到所有成分切分并标记完成，输出未被标记的成分名称，标记为(-*-)；其中，设t为x的句子成分，freq(x,t)表示字符串x及对应的句子成分t在训练树库中出现的次数；设成分序列为X1,X2,X3…,Xn，将标记的成分与训练树库成分字段比对，自动生成待校树库；获取待校树库节点；

成分搭配检测步骤，遍历扫描待校树库中X1,X2,X3…Xn成分，并与训练树库成分字段匹配，具体方法包括：

步骤41，查找待校树库的根节点；

步骤42，访问该节点；

步骤43，判断该节点是否有未访问的子节点，如果有，执行步骤44；如果没有，执行步骤45；

步骤44，访问最左侧未被访问的子节点，并将该节点与根节点组合搭配，与训练树库成分字段比对，如果正确，则输出该节点对应成分，执行步骤42；如果错误，则输出该节点对应成分并标记为(-*-)，执行步骤42；

步骤45：判断该节点是否为根结点，如果是，执行步骤46；如果不是，执行步骤47；

步骤46：将该节点与训练树库成分字段比对，如果正确，则输出该节点对应成分；如果错误，则输出该节点对应成分并标记为(-*-)；

步骤47：返回该节点的父节点，执行步骤43。

5.根据权利要求1所述的一种基于规则和语料库的汉语语病自动检测方法，其特征在于，语病检测包括语义表达检测，所述语义表达检测包括以下步骤：

语义表达分词结果获取步骤，获取待校文本句子分层结果及自动分词步骤的待校文本分词结果，所述分词结果为自动分词步骤依次输出的固定式、词语和单字词，并按固定式、词语和单字词在单句中排列顺序依次标记位置为1,2,3,...,Z；

语义成分提取步骤，在分词基础上，从左往右依次切分出长度不等的字符串，即从第一个词开始，依次切分出一个词、两个词、三个词、…、Z个词的字符串；计算切分出的字符串x为语义成分的概率$\tilde{P}(t|x)=\frac{freq(x,t)}{\sum_{x,t}freq(x,t)}$，如果$\tilde{P}(t|x)\geq 0.2$，则切分出字符串x，并标记语义成分名称；如果$\tilde{P}(t|x)<0.2$，表示该字符串x不是句中语义成分，则判断下一个字符串，直到所有成分切分并标记完成，输出未被标记的成分名称，标记为(-*-)；其中，设t为x的语义成分，freq(x,t)表示字符串x及对应的语义成分t在预先构建的语义训练树库中出现的次数；设语义成分序列为X1,X2,X3…,Xn，将标记的语义成分与语义训练树库成分字段比对，自动生成待校树库；获取待校树库节点；

语义成分搭配检测步骤，利用语义训练树库及其规则字段，遍历执行X1,X2,X3…Xn语义搭配检测，具体方法包括：

步骤51，查找待校树库的根节点；

步骤52,访问该节点;

步骤53,判断该节点是否有未访问的子节点,如果有,执行步骤54;如果没有,执行步骤55;

步骤54,访问最左侧未被访问的子节点,并将该节点与根节点组合搭配,与语义训练树库成分字段和规则字段比对,如果正确,则输出该节点对应语义成分,执行步骤52;如果错误,则输出该节点对应语义成分并标记为(-*-),执行步骤52;

步骤55,判断该节点是否为根结点,如果是,执行步骤56;如果不是,执行步骤57;

步骤56,将该节点与语义训练树库成分字段和规则字段比对,如果正确,则输出该节点对应成分;如果错误,则输出该节点对应成分并标记为(-*-);

步骤57,返回该节点的父节点,执行步骤53。

6.一种基于规则和语料库的汉语语病自动检测设备,其特征在于,包括:

文本获取装置,用于获取待校文本数据;

句子分层装置,用于读取文本,获取文本中的句子数量,并将获取的文本划分为单句;

正向切分字符串装置,以单句为单位,获取单句字符串长度,从左往右依次切分出长度不等的字符串,即从第一个字符开始,依次切分出N个字符、N-1个字符、N-2个字符、…、两个字符、一个字符的字符串,N为单句字符串长度;

逆向切分字符串装置,以单句为单位,获取单句字符串长度,从右往左依次切分出长度不等的字符串,即从最后一个字符开始,依次切分出一个字符、两个字符、三个字符、…、N个字符的字符串;

自动分词装置,用于将切分出的字符串依次与预先构建的语料库中的固定式语料库和词语语料库比对,若匹配成功,则输出该字符串并标记序列号为1,2,3,...,z,若匹配失败,则将未被标记序列号的其他单个字符逐一与预先构建的语料库中的单字词语料库比对,若匹配成功,则为单字词,输出该单字词并标记对应的序列号为z+1,z+2,z+3,...,若匹配失败,则与预先构建的语料库中的非单字词语料库比对,若匹配成功,则为非单字词,输出该非单字词,若匹配失败,则保留该字符;

语病检测装置,用于根据切分字符串装置和自动分词装置的结果及预先构建的语料库进行语病检测;

所述预先构建的语料库包括固定式语料库、字母语料库、标点语料库、拼音语料库、繁简字语料库、语句语料库、词语语料库和单字词语料库;所述语病检测包括错别字检测、用词不当检测、句法结构检测和语义表达检测。

7.根据权利要求6所述的一种基于规则和语料库的汉语语病自动检测设备,其特征在于,语病检测装置包括用词不当检测装置,所述用词不当检测装置包括:

用词不当分词结果获取子装置,用于获取待校文本句子分层结果及自动分词步骤的待校文本分词结果,所述分词结果为自动分词步骤依次输出的固定式、词语和单字词,并按固定式、词语和单字词在单句中排列顺序依次标记位置为1,2,3,...,Z;

用词不当检测子装置,用于进行用词不当检测,所述用词不当检测包括以下步骤:

相邻词位置检测子步骤,结合预先构建的语料库中的词语语料库和单字词语料库中位置字段的标记,将待校文本中的词从自由与粘着两个角度自动标注位置信息;判断待校文本是否有定位词的标记,将待校文本中的定位词和相邻词与词语语料库和单字词语料库位

置搭配规则对比，判断是否正确，如果错误，则输出错误词并标记为(*)；

相邻词词性检测子步骤，结合预先构建的语料库中的词语语料库和单字词语料库的词性字段的标记，将待校文本中的词自动标注词性信息；将待校文本中相邻词的词性与词语语料库和单字词语料库中的词性搭配规则进行匹配，判断待校文本中相邻词之间的词性是否能搭配，如果不能搭配，则输出错误词并标记为(*)；

相邻词语义搭配检测子步骤，结合预先构建的语料库中的词语语料库和单字词语料库的语义字段的标记，将待校文本中的词自动标注语义信息；将待校文本中相邻词的语义与词语语料库和单字词语料库中的语义搭配规则进行匹配，判断待校文本中相邻词之间的语义是否能搭配，如果不能搭配，则输出错误词并标记为(*)；

不相邻词语义搭配检测子步骤，采用互信息算法，对i和j这两个词，通过对比语句语料库，计算它们的互信息值$Q(i,j)=log_2\frac{P(i,j)}{P(i)P(j)}$，用Q(i,j)的大小判断这两个词语义的组合情况，当Q(i,j)>0时，表示i和j这两个词语义组合正确，输出i和j；当Q(i,j)＝0时，表示i和j语义组合不明确，则输出i和j并标注为(?)；当Q(i,j)<0时，表示i和j语义组合不正确，则输出i和j并标注为(*)；其中，i和j分别为待校文本单句中任意两个词，P(i,j)为i和j这两个词共现的频率，P(i)和P(j)分别为i和j这两个词出现的频率。

8.根据权利要求6所述的一种基于规则和语料库的汉语语病自动检测设备，其特征在于，语病检测装置包括句法结构检测装置，所述句法结构检测装置包括：

句法结构分词结果获取子装置，用于获取待校文本句子分层结果及自动分词步骤的待校文本分词结果，所述分词结果为自动分词步骤依次输出的固定式、词语和单字词，并按固定式、词语和单字词在单句中排列顺序依次标记位置为$1，2，3，...，Z$；

标记虚词子装置，用于标记待校文本单句中虚词，与单字词语料库位置字段比对，自动标注虚词位置；

句法成分提取子装置，用于在分词基础上，从左往右依次切分出长度不等的字符串，即从第一个词开始，依次切分出一个词、两个词、三个词、…、Z个词的字符串；计算切分出的字符串x为句子成分的概率$\tilde{P}(t|x)=\frac{freq(x,t)}{\sum_{x,t}freq(x,t)}$，如果$\tilde{P}(t|x)\geq 0.2$，则切分出字符串x，并标记成分名称；如果$\tilde{P}(t|x)<0.2$，表示该字符串x不是句中句子成分，则判断下一个字符串，直到所有成分切分并标记完成，输出未被标记的成分名称，标记为(-*-)；其中，设t为x的句子成分，freq(x,t)表示字符串x及对应的句子成分t在训练树库中出现的次数；设成分序列为X1,X2,X3…,Xn，将标记的成分与训练树库成分字段比对，自动生成待校树库；获取待校树库节点；

成分搭配检测子装置，用于遍历扫描待校树库中X1,X2,X3…Xn成分，并与训练树库成分字段匹配，具体方法包括：

步骤41，查找待校树库的根节点；

步骤42，访问该节点；

步骤43，判断该节点是否有未访问的子节点，如果有，执行步骤44；如果没有，执行步骤45；

步骤44,访问最左侧未被访问的子节点,并将该节点与根节点组合搭配,与训练树库成分字段比对,如果正确,则输出该节点对应成分,执行步骤42;如果错误,则输出该节点对应成分并标记为(-*-),执行步骤42;

步骤45:判断该节点是否为根结点,如果是,执行步骤46;如果不是,执行步骤47;

步骤46:将该节点与训练树库成分字段比对,如果正确,则输出该节点对应成分;如果错误,则输出该节点对应成分并标记为(-*-);

步骤47:返回该节点的父节点,执行步骤43。

9.根据权利要求6所述的一种基于规则和语料库的汉语语病自动检测设备,其特征在于,语病检测装置包括语义表达检测装置,所述语义表达检测装置包括:

语义表达分词结果获取子装置,用于获取待校文本句子分层结果及自动分词步骤的待校文本分词结果,所述分词结果为自动分词步骤依次输出的固定式、词语和单字词,并按固定式、词语和单字词在单句中排列顺序依次标记位置为1,2,3,...,Z;

语义成分提取子装置,用于在分词基础上,从左往右依次切分出长度不等的字符串,即从第一个词开始,依次切分出一个词、两个词、三个词、…、Z个词的字符串;计算切分出的字符串x为语义成分的概率$\tilde{P}(t|x)=\frac{freq(x,t)}{\sum_{x,t}freq(x,t)}$,如果$\tilde{P}(t|x)\geq 0.2$,则切分出字符串x,并标记语义成分名称;如果$\tilde{P}(t|x)<0.2$,表示该字符串x不是句中语义成分,则判断下一个字符串,直到所有成分切分并标记完成,输出未被标记的成分名称,标记为(-*-);其中,设t为x的语义成分,freq(x,t)表示字符串x及对应的语义成分t在预先构建的语义训练树库中出现的次数;设语义成分序列为X1,X2,X3…,Xn,将标记的语义成分与语义训练树库成分字段比对,自动生成待校树库;获取待校树库节点;

语义成分搭配检测子装置,用于利用语义训练树库及其规则字段,遍历执行X1,X2,X3…Xn语义搭配检测,具体方法包括:

步骤51,查找待校树库的根节点;

步骤52,访问该节点;

步骤53,判断该节点是否有未访问的子节点,如果有,执行步骤54;如果没有,执行步骤55;

步骤54,访问最左侧未被访问的子节点,并将该节点与根节点组合搭配,与语义训练树库成分字段和规则字段比对,如果正确,则输出该节点对应语义成分,执行步骤52;如果错误,则输出该节点对应语义成分并标记为(-*-),执行步骤52;

步骤55,判断该节点是否为根结点,如果是,执行步骤56;如果不是,执行步骤57;

步骤56,将该节点与语义训练树库成分字段和规则字段比对,如果正确,则输出该节点对应成分;如果错误,则输出该节点对应成分并标记为(-*-);

步骤57,返回该节点的父节点,执行步骤53。

10.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现权利要求1至5中任一项所述的方法的步骤。

# 基于规则和语料库的汉语语病自动检测方法及设备

## 技术领域

[0001] 本发明涉及信息处理技术领域，尤其涉及一种基于规则和语料库的汉语语病自动检测方法及设备。

## 背景技术

[0002] 汉语语病自动检测相对英文、日文等检测研究来说，起步较晚，缘于技术和汉语自身的特点，中文文本自动检测发展较慢。面对海量的待校对文本信息，汉语语病自动检测是亟待有效解决的一大难题。

[0003] 目前，已有文献开始提出文本校对的思路和方法。从现有研究对象和进程看，中文字词检测已逐渐发展起来，错别字自动检测的理论研究和应用研究都取得了一定成效，但针对汉语语病的检测却鲜有人提及。至今，校对研究常限于某一领域(如，基于形态学、基于相邻词性的连接规则、基于某一语法规则或某一语义规则)设计相应方法，具体而言，目前比较成熟的校对方法主要有两类：

[0004] 基于特征的校对方法(包括词法特征、句法特征)。其方法是对词句进行分类，分析词与词或单一句法成分的搭配。该方法能解决一部分典型案例，但未能找到合适的切入点，未从整体角度分析汉语的特点，未综合分析汉语内部各要素之间的组合聚合搭配规则，进而依次成系统地进行检测。

[0005] 基于语义的校对方法(大多使用SUM算法、决策树、Bayes算法)。该检测方法结合了汉语的特征，符合汉语内部组合规律。但设计方法时直接从语义搭配开始，在极有限的样本分析基础上设计算法，存在样本信息有限、检测类型单一、参数需要不断调整、代表性不强等方面的不足。

[0006] 从现有研究来看，我们面临的主要问题有：如何深入汉语本体研究，进一步分析和探讨汉语内部各构成要素之间的关系和规律；如何将汉语本体研究与信息处理技术充分结合，将不同学科不同领域的知识融会贯通；如何科学全面地设计符合汉语内部规律的语病检测方法等。回到汉语本体研究层面，我们结合语言学研究再做新的审核。索绪尔《普通语言学教程》(2009)中分析了汉语的运行特点和规律：汉语中的字词是线性的，它们彼此结成以线条性为基础的关系，单独成为一个要素，这些要素又按照一定的规则一个挨着一个进行排列组合。陆俭明《现代汉语语法研究教程》(2005)认为汉语语法不仅是构成关系，也是组合关系，它内部的规则，就是指小的结合体组成大的结合体所依据的一系列规则。结合众多语言学家的研究成果，汉语语病的自动检测，还是得从组合构成着手，以字词组合为基础，检测句子成分之间的搭配及相互关系，再上升到语义和语用分析。也就是说，汉语是成系统的，应系统地分析字词组合与搭配、句子结构、内部语义等，不能割裂它们之间的联系。

## 发明内容

[0007] 本发明所要解决的技术问题是：针对现有技术存在的问题，本发明提供一种基于规则和语料库的汉语语病自动检测方法及设备，从词法、句法、语义等角度进行检测，自动

检测文本中的各类语病问题。

[0008] 本发明提供的一种基于规则和语料库的汉语语病自动检测方法，包括文本获取、句子分层、自动分词和语病检测；所述文本获取为获取待校文本数据；所述句子分层包括读取文本，获取文本中的句子数量，并将获取的文本划分为单句；所述自动分词包括以下步骤：正向切分字符串步骤，以单句为单位，获取单句字符串长度，从左往右依次切分出长度不等的字符串，即从第一个字符开始，依次切分出N个字符(第一个字符至第N个字符)、N-1个字符(第一个字符至第N-1个字符、第二个字符至第N个字符)、N-2个字符(第一个字符至第N-2个字符、第二个字符至第N-1个字符、第三个字符至第N个字符)、…、两个字符(如，第一个字符和第二个字符、第二个字符和第三个字符、…、第N-1个字符和第N个字符)、一个字符的字符串；逆向切分字符串步骤，以单句为单位，获取单句字符串长度，从右往左依次切分出长度不等的字符串，即从最后一个字符开始，依次切分出一个字符、两个字符(如，第N个字符和第N-1个字符、第N-1个字符和第N-2个字符、…、第2个字符和第1个字符)、三个字符(如，第N个字符至第N-2个字符、第N-1个字符至第N-3个字符、…、第3个字符至第1个字符)、…、N个字符(第N个字符至第1个字符)的字符串，N为单句字符串长度；自动分词步骤，将切分出的字符串依次与预先构建的语料库中的固定式语料库和词语语料库比对，若匹配成功，则输出该字符串并标记序列号(1,2,3,...,z)，若匹配失败，则将单句中落单的字符(未被标记序列号的其他单个字符)逐一与预先构建的语料库中的单字词语料库比对，若匹配成功，则为单字词，输出该单字词并标记对应的序列号(z+1,z+2,z+3,...)，若匹配失败，则与预先构建的语料库中的非单字词语料库比对，若匹配成功，则为非单字词，输出该非单字词，若匹配失败，则保留该字符；所述语病检测为根据所述自动分词的结果和预先构建的语料库进行语病检测。

[0009] 进一步，所述语病检测包括错别字检测、用词不当检测、句法结构检测、语义表达检测，语病检测可以包括这四种检测中的一种或几种。

[0010] 本发明另一方面还提供一种基于规则和语料库的汉语语病自动检测设备，包括：

[0011] 文本获取装置，用于获取待校文本数据；句子分层装置，用于读取文本，获取文本中的句子数量，并将获取的文本划分为单句；正向切分字符串装置，用于以单句为单位，获取单句字符串长度，从左往右依次切分出长度不等的字符串，即从第一个字符开始，依次切分出N个字符(第一个字符至第N个字符)、N-1个字符(第一个字符至第N-1个字符、第二个字符至第N个字符)、N-2个字符(第一个字符至第N-2个字符、第二个字符至第N-1个字符、第三个字符至第N个字符)、…、两个字符(如，第一个字符和第二个字符、第二个字符和第三个字符、…、第N-1个字符和第N个字符)、一个字符的字符串；逆向切分字符串装置，用于以单句为单位，获取单句字符串长度，从右往左依次切分出长度不等的字符串，即从最后一个字符开始，依次切分出一个字符、两个字符(如，第N个字符和第N-1个字符、第N-1个字符和第N-2个字符、…、第2个字符和第1个字符)、三个字符(如，第N个字符至第N-2个字符、第N-1个字符至第N-3个字符、…、第3个字符至第1个字符)、…、N个字符(第N个字符至第1个字符)的字符串，N为单句字符串长度；自动分词装置，用于将切分出的字符串依次与预先构建的语料库中的固定式语料库和词语语料库比对，若匹配成功，则输出该字符串并标记序列号(1,2,3,...,z)，若匹配失败，则将单句中落单的字符(未被标记序列号的其他单个字符)逐一与预先构建的语料库中的单字词语料库比对，若匹配成功，则为单字词，输出该单字词并标记

对应的序列号(z+1,z+2,z+3,...),若匹配失败,则与预先构建的语料库中的非单字词语料库比对,若匹配成功,则为非单字词,输出该非单字词,若匹配失败,则保留该字符;语病检测装置,用于根据所述切分字符串装置和自动分词装置的结果及预先构建的语料库进行语病检测。

[0012] 进一步,所述语病检测装置包括错别字检测装置、用词不当检测装置、句法结构检测装置、语义表达检测装置,语病检测装置可以包括这四种检测装置中的一种或几种。

[0013] 本发明另一方面还提供一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如上所述的方法的步骤。

[0014] 随着网络电子文本数量的增加,语病出现的概率也将逐渐增大,据统计,人们在互联网上每天传输的数据超过了整个19世纪的全部数据的总和,面对海量的文字信息,这是人工校对所不能应对的,这就需要我们研究出自动检测语病的方法和设备。本发明避免过去研究中仅从某个角度进行探讨的思路,综合考虑了语病的类型及出现的原因,且充分分析汉语语病规律和特点,从根源出发,科学地结合语言学知识设计检测角度和内容,将弥补已有研究中无充分语言学理论指导的空白。建立相应语料库和结构树库、语义树库,结合了信息处理中的思路,设计了基于规则和语料库的汉语语病检测方法,在理论上更为可靠。

[0015] 本发明设计了一种全面检测的方法,将汉语语病整个系统纳入研究之中,环环相扣。且前景广阔,可以通用于各个与汉字录入相关的电子设备中,它将不仅可以解决因输入法因素造成的语病问题,还可以进一步检测出手写体文字中的语病,可以通过对手写体进行识别、对图片中的文字进行扫描匹配、对人工录入的文字、语音录入等都能进行检测。

## 附图说明

[0016] 本发明将通过举例并参照附图的方式说明,其中:

[0017] 图1为本发明实施例的汉语语病检测流程图;

[0018] 图2为本发明实施例的字处理方法流程图;

[0019] 图3为本发明实施例的词处理方法流程图;

[0020] 图4为本发明实施例的词处理的具体实施例流程图;

[0021] 图5为本发明实施例的句子类型示意图;

[0022] 图6为本发明实施例的句法成分训练树库标记示意图;

[0023] 图7为本发明实施例的结构处理方法流程图;

[0024] 图8为本发明实施例的结构处理中的搭配检测流程图;

[0025] 图9和图10为本发明实施例的句法语义关系示意图;

[0026] 图11为本发明实施例的语义成分训练树库标记示意图;

[0027] 图12为本发明实施例的语义处理方法流程图;

[0028] 图13为本发明实施例的语义处理中的语义搭配检测流程图。

## 具体实施方式

[0029] 本说明书中公开的所有特征,或公开的所有方法或过程中的步骤,除了互相排斥的特征和/或步骤以外,均可以以任何方式组合。

[0030] 本说明书中公开的任一特征,除非特别叙述,均可被其他等效或具有类似目的的

替代特征加以替换。即，除非特别叙述，每个特征只是一系列等效或类似特征中的一个例子而已。

[0031] 本发明基于汉语内部构成和组合规律，分析汉语结构和语义要素，重分语病类型，构建多重语料库，并设计检测方法。

[0032] 本发明测试过程主要基于oracle数据库和Myeclipse软件，技术方案及措施如下：

[0033] 1.语病类型重组和前期处理过程

[0034] 根据语病划分标准，汉语语病主要有三种分类体系，为更好地进行检测，我们重组现有划分类型，将其分为四类，依次为用字、用词、结构、语义表达错误。

[0035] 前期处理过程包括文本获取、句子分层和自动分词。

[0036] 所述文本获取为获取待校文本数据。

[0037] 所述句子分层包括读取文本，获取文本中的句子数量(查找文本中句号、感叹号、问号、分号、段末省略号数量，此处，每个符号表示一个句子)，并将获取的文本划分为单句。

[0038] 所述自动分词包括以下步骤：

[0039] 切分字符串步骤，以单句为单位，获取单句字符串长度(即字符个数)，赋给变量N(N>0)，设每个单句中的字符序列号为1,2,3,...N。循环执行以下操作(直到该单句所有字符切分完毕)：

[0040] 正向切分(从左往右，从句子到单个字符)字符串步骤，从左往右依次切分出长度不等的字符串，即从第一个字符开始依次切分出N个字符(第一个字符至第N个字符)、N-1个字符(第一个字符至第N-1个字符、第二个字符至第N个字符)、N-2个字符(第一个字符至第N-2个字符、第二个字符至第N-1个字符、第三个字符至第N个字符)、…、两个字符(如，第一个字符和第二个字符、第二个字符和第三个字符、…、第N-1个字符和第N个字符)、一个字符的字符串；

[0041] 逆向切分(从右往左，从单个字符到句子)字符串步骤，以单句为单位，获取单句字符串长度，从右往左依次切分出长度不等的字符串，即从最后一个字符开始，依次切分出一个字符、两个字符(如，第N个字符和第N-1个字符、第N-1个字符和第N-2个字符、…、第2个字符和第1个字符)、三个字符(如，第N个字符至第N-2个字符、第N-1个字符至第N-3个字符、…、第3个字符至第1个字符)、…、N个字符(第N个字符至第1个字符)的字符串；

[0042] 自动分词步骤，将切分出的字符串与预先构建的语料库中的固定式语料库和词语语料库比对，若匹配成功，则输出该字符串并标记序列号(1,2,3,...,z(z>＝0))，若匹配失败，则将单句中落单的字符(未被标记序列号的其他单个字符)逐一与预先构建的语料库中的单字词语料库比对，若匹配成功，则为单字词，输出该单字词并标记对应的序列号(z+1,z+2,z+3,...)，若匹配失败，则与预先构建的语料库中的非单字词语料库比对，若匹配成功，则为非单字词，输出该非单字词，若匹配失败，则保留该字符。

[0043] 任意选取语句语料库中50条语料(50个单句，共786个词)进行自动分词测试，发现正向切分字符串(单个字符串切分到整句切分)进行分词的准确率为90.1％，正向切分字符串(整句切分到单个字符串切分)进行分词的准确率为93.1％；

[0044] 逆向切分字符串(整句切分到单个字符串切分)进行分词的准确率为93.1％，与正向切分(单个字符串切分到整句切分)结合的准确率为95％，与正向切分(整句切分到单个字符串切分)结合的准确率为93.1％；

[0045] 逆向切分字符串(单个字符串切分到整句切分)进行分词的准确率为96.4%,与正向切分(单个字符串切分到整句切分)结合的准确率为96.8%,与正向切分(整句切分到单个字符串切分)结合的准确率为98.9%,所以将正向切分(整句切分到单个字符串切分)和逆向切分(单个字符串切分到整句切分)结合起来。

[0046] 对待校文本进行前期处理后,即可进行语病检测,语病检测为根据自动分词的结果和预先构建的语料库进行语病检测。语病检测包括错别字检测、用词不当检测、句法结构检测、语义表达检测,语病检测可以包括这四种检测中的一种或几种。且在一些实施例中,语病检测包括的几种检测可以并列执行,而在其他一些实施例中,这几种检测可以依次执行,且这几种检测的前后顺序可以根据情况进行选择。为更加清楚地对本发明进行说明,本发明实施例中根据错别字检测、用词不当检测、句法结构检测和语义表达检测依次执行的步骤进行详细说明,如图1所示。

[0047] 2.错别字检测

[0048] 文本中用字错误主要指错别字、繁体字(视情况而定)和不规范字,其中以错别字为主(也有人将字级错误统称为错别字)。检测文本中的用字错误,我们将通过以下步骤实现:

[0049] 2.1构建多重语料库

[0050] 2.1.1建立固定式语料库

[0051] 在汉语书面表达中,字词通过组合会形成固定式结构(如,成语、熟语、专业术语等)和非固定式结构(固定式结构以外的句子成分、词语、单字等)。

[0052] 2.1.1.1创建一个新的语料库,命名为固定式语料库,建七个字段,分别命名为固定式、位置、词性、语义、位置搭配规则、词性搭配规则、语义搭配规则。

[0053] 2.1.1.2将字词典中收录的成语、熟语、谚语、歇后语、专业术语、人名构成、地名构成、百分数形式、小数形式、数目字、字母、诗词、文言、名句名篇、被收录词典的方言词语、简称、重叠词、音译词等录入固定式语料库(固定式字段)中。

[0054] 2.1.2建立非固定式语料库

[0055] 非固定式结构主要由单字词(即能单独成词的字,也叫成词单字)、非单字词(即不能单独成词的字,也叫不成词单字)和词语(非单音节词语,即两个及以上音节组成的词)构成。

[0056] 2.1.2.1创建三个新的语料库,依次命名为单字词语料库、非单字词语料库、词语语料库,分别建单字词字段、非单字词字段、词语字段。

[0057] 2.1.2.2词语语料库、单字词语料库增设位置字段、词性字段、语义字段、位置搭配规则字段、词性搭配规则字段、语义搭配规则字段。

[0058] 2.1.2.3将字词典中收录的单字词、非单字词、词语依次录入单字词语料库(单字词字段)、非单字词语料库(非单字词字段)、词语语料库(词语字段)中。

[0059] 2.1.3建立繁简字对应语料库

[0060] 2.1.3.1创建一个新的语料库,命名为繁简字语料库,建繁体字字段、简体字字段。

[0061] 2.1.3.2参照《繁简字对应表》,将繁简字一一录入繁简字语料库(繁体字字段、简体字字段)中。

[0062] 2.1.4建立汉语拼音语料库

[0063] 2.1.4.1创建一个新的语料库,命名为拼音语料库,建单字字段、拼音字段。

[0064] 2.1.4.2将《汉语大字典》收录的单字和它对应的拼音录入拼音语料库(单字字段、拼音字段)中。

[0065] 2.1.5建立标点语料库

[0066] 2.1.5.1创建一个新的语料库,命名为标点语料库,建标点字段。

[0067] 2.1.5.2将汉语中所有标点和其他的符号录入标点语料库(标点字段)中。

[0068] 2.1.6建立数字、字母语料库

[0069] 2.1.6.1创建一个新的语料库,命名为字母语料库,建字母字段。

[0070] 2.1.6.2将数字0-9及26个字母(大写、小写)、英语单词(《牛津高阶英汉双解词典》电子版)录入字母语料库(字母字段)中。

[0071] 2.1.7建立语病语料库

[0072] 2.1.7.1创建一个新的语料库,命名为语病语料库,建错误字段、正确字段。

[0073] 2.1.7.2搜集电子刊物、网页(如,百度百科、360百科等)、文本中的病句(包括用字、用词、结构、语义错误),将155万字的语病语料录入语病语料库(错误字段、正确字段)中。

[0074] 2.1.8建立语句语料库

[0075] 2.1.8.1创建一个新的语料库,命名为语句语料库,建语句字段。

[0076] 2.1.8.2搜集电子刊物、文学著作、学科论文等文本语料(包括古代汉语语料、现代汉语语料),以句子为单位,将一亿三千多万字语料录入语句语料库(语句字段)中。

[0077] 2.2错别字检测方法设计,如图2所示。

[0078] 随机选出语病语料库中100条错别字语料,并进行测试,正向检测召回率(检测出的语病总数/训练语料中的语病总数)为0.89,准确率(检测正确的总数/检测出的语病总数)为0.8,逆向检测召回率为0.91,准确率为0.88,正向与逆向结合的召回率为0.97,准确率达到0.95,所以将正向和逆向检测结合起来。

[0079] 2.2.1正向检测

[0080] 循环执行以下操作,直到所有单句中的字符检测完毕:

[0081] 2.2.1.1字母检测。

[0082] 2.2.1.1.1判断切分出的字符串是否有数字和(或)字母,若是,则执行2.2.1.1.2;若不是,则执行2.2.1.2。

[0083] 2.2.1.1.2将切分出的字符串与字母语料库比对。如果形式正确,则输出该字符串,并执行2.2.1.2;如果形式错误,则输出该字符串并标记为(*),执行2.2.1.2。

[0084] 2.2.1.2标点检测。

[0085] 2.2.1.2.1判断切分出的字符串中是否含有标点符号或特殊符号,若是,则执行2.2.1.2.2;若不是,则执行2.2.1.3。

[0086] 2.2.1.2.2将切分出的字符串与标点语料库比对,如果形式正确,则输出该字符串,执行2.2.1.3;如果形式错误(如,乱码、无意义的符号),则输出该字符串并标记为(*),执行2.2.1.3。

[0087] 2.2.1.3拼音检测。

[0088] 2.2.1.3.1判断切分出的字符串是否有拼音(查找文本中是否包含字母和声调,或

者纯拼音字母)，若是，执行2.2.1.3.2；若不是，执行2.2.1.4。

[0089] 2.2.1.3.2将切分出的字符串与拼音语料库(单字字段、拼音字段)比对，如果形式正确，则输出该字符串，执行2.2.1.4；如果形式错误，输出该字符串并标记为(*)，执行2.2.1.4。2.2.1.4繁体字检测。

[0090] 2.2.1.4.1将待校文本与繁简字语料库比对，判断切分出的字符串是否有繁体字，若是，则执行2.2.1.4.2；若不是，执行2.2.1.5。

[0091] 2.2.1.4.2获取繁体字数量，赋给变量E(E>＝0)。设繁体字序列号为1,2,3,...E，循环执行2.2.1.4.3。

[0092] 2.2.1.4.3将序列号为1,2,3,...E的繁体字逐一提取，判断它是否属于引用或特别使用情况(一般位于引号、冒号、书名号、括号内)，若是，则输出该繁体字并执行2.2.1.5；若不是，输出该繁体字，标记为(*)。

[0093] 2.2.1.5将自动分词步骤判断为单字词的单字与下一单字组合，与语句语料库比对，若匹配成功，则输出该单字；若匹配失败，则执行3.3。

[0094] 2.2.1.6将自动分词步骤判断为非单字词的单字与下一单字组合，与语句语料库比对，若匹配成功，则输出该单字；若匹配失败，则输出该单字并标记为(*)。

[0095] 2.2.2逆向检测

[0096] 重复执行2.2.1.5—2.2.1.6环节，以单句为单位，从右至左，将自动分词步骤判断为单字词的单字与下一单字组合，与语句语料库比对，若匹配成功，则输出该单字；若匹配失败，则执行3.3；将自动分词步骤判断为非单字词的单字与下一单字组合，与语句语料库比对，若匹配成功，则输出该单字；若匹配失败，则输出该单字并标记为(*)。

[0097] 3.用词不当检测

[0098] 文本中的用词错误包括词语使用不当和生造词语，主要检测词语、单字词和部分固定式结构的使用和搭配。根据汉语词语组合特征和规律，文本中对用词的检测需要通过位置、词性和语义搭配来判断。对此，将词语语料库(词语字段)、单字词语料库(单字词字段)和固定式语料库(固定式字段)中的词语和单字词一一描写和标记其搭配位置、词性和语义。

[0099] 3.1描写和标记

[0100] 3.1.1标记词(特指词语和单字词，下同)的位置信息

[0101] 3.1.1.1将词语语料库和单字词语料库(固定式结构均为不定位词组)中的词语和单字词按定位与不定位特征进行分类。

[0102] 3.1.1.2标记定位词的位置信息：前接成分(处在某个词的后面做后缀)标记为“h”，后接成分(处在某个词的前面做前缀)标记为“k”，录入位置字段中。参照语句语料库中的语料，分别描写定位词能搭配的词及所处位置，录入位置搭配规则字段中。

[0103] 3.1.1.3将不定位词语归类，不做标记。

[0104] 3.1.2标记词的词性

[0105] 3.1.2.1根据《汉语大词典》《汉语大字典》《现代汉语词典》等工具书收录词条的词性信息，逐个标记词语语料库、单字词语料库中词语和单字词的词性(固定式结构由词组合构成，统一标记为“i”)，录入词性字段中。以下为词性标注的名称及对应符号：

[0106] 表1

[0107]

| 词性标注 | | | | | | | |
|---|---|---|---|---|---|---|---|
| 名称 | 符号 | 名称 | 符号 | 名称 | 符号 | 名称 | 符号 |
| 普通名词 | n | 一般动词 | v | 量词 | q | 助词 | u |
| 时间名词 | nt | 及物动词 | vt | 副词 | d | 叹词 | e |
| 方位名词 | nd | 不及物动词 | vi | 程度副词 | cd | 拟声词 | o |
| 处所名词 | nl | 形容词 | a | 否定副词 | fd | 代词 | r |
| 地名 | ns | 性质形容词 | xa | 介词 | p | 固定结构 | i |
| 机构名 | ni | 状态形容词 | za | 连词 | c | 前接成分 | h |
| 人名 | nh | 区别词 | f | 数词 | m | 后接成分 | k |
| 错别字词 | (*) | 疑似错误 | (？) | 错误结构表达 | (-*-) | 其它字符 | X |

[0108] 3.1.2.2标记词能搭配的词性信息。

[0109] 根据汉语字词之间词性搭配的规则，逐个描写词语语料库、单字词语料库中的词语和单字词能搭配的词性，录入词性搭配规则字段中。汉语中有的词语、单字词虽然词性相同，但词性搭配情况不同，需要一一标记出来。如，“红”和“通红”都是形容词，但否定副词、程度副词只能修饰前者，不能修饰后者。再如，副词一般不修饰名词(除了极特殊的情况)，而这种特殊情况只能参照语句语料库中语料，标记时将能修饰名词的副词和搭配的情况一一描写出来，录入词性搭配规则字段中。

[0110] 统计发现，汉语单字词大多在虚词和代词中，所以，对单字词的词性描写有一定量的限定。

[0111] 3.1.3标记词的语义特征

[0112] 3.1.3.1根据《汉语大词典》《汉语大字典》《现代汉语词典》等工具书收录词条的语义信息，逐个标记词语、单字词、固定式结构的语义特征(指某个词或词组所特有的、能对其所在的句法格式起制约作用的，并足以区别于其他小类实词的语义要素)，分别录入词语语料库、单字词语料库、固定式语料库的语义字段中。

[0113] 3.1.3.2对词语、单字词、固定式结构前后可能出现的词通过语义指向(指句法结构的某一成分在语义上和其他成分相匹配的可能性。如，动词可以根据最多能够搭配的名词数量来判断它的语义指向)来描写，分别录入词语语料库、单字词语料库、固定式语料库的语义搭配规则字段中。

[0114] 3.2根据前期处理过程中的句子分层，获取待校文本分句结果，根据前期处理过程中的自动分词步骤，获取待校文本分词结果(依次输出的固定式、词语、单字词的序列)，并按固定式、词语和单字词在单句中排列顺序依次标记位置为1,2,3,...,Z(Z>0)。具体流程如图3所示。

[0115] 3.3检测方法

[0116] 3.3.1设I(I＝1;I<＝Z-1)和J(J＝I+1;J<＝Z)分别表示待校文本中序列号为1,2,3…Z所对应的相邻两个词,I和J循环递增。

[0117] 3.3.2相邻词位置检测

[0118] 3.3.2.1结合词语语料库、单字词语料库中位置字段的标记,将待校文本中的词从自由与粘着(或定位与不定位)两个角度自动标注位置信息。

[0119] 3.3.2.2判断待校文本是否有定位词(即获取3.3.2.1标记的待校文本中是否有“h”“k”符号)的标记。

[0120] 3.3.2.2.1将待校文本中的定位词和相邻词与词语语料库、单字词语料库位置搭配规则对比,判断是否正确。

[0121] 3.3.2.2.2如果错误,输出错误词,并标注为(*)。

[0122] 3.3.2.2.3如果正确,执行3.3.3。

[0123] 3.3.3相邻词词性检测

[0124] 结合词语语料库、单字词语料库的词性字段的标记,将待校文本中的词自动标注词性信息。

[0125] 3.3.3.1循环I和J,将待校文本中相邻词的词性与词语语料库、单字词语料库中的词性搭配规则字段进行匹配,判断待校文本中相邻词之间的词性是否能搭配。

[0126] 3.3.3.2如果能搭配,执行3.3.4。

[0127] 3.3.3.3如果不能搭配,输出错误词,并标注为(*)。

[0128] 3.3.4相邻词的语义搭配检测

[0129] 结合词语语料库、单字词语料库的语义字段的标记,将待校文本中的词自动标注语义信息。

[0130] 3.3.4.1循环I和J,将待校文本中相邻词的语义与词语语料库、单字词语料库中的语义搭配规则字段进行匹配,判断待校文本中相邻词之间的语义是否能搭配。

[0131] 3.3.4.2如果能搭配,执行3.3.5。

[0132] 3.3.4.3如果不能搭配,输出错误词,并标注为(*)。

[0133] 3.3.5不相邻词的语义搭配检测

[0134] 汉语中判断一个句子用词是否恰当,从语义角度考虑,不仅要检测其相邻词的语义搭配,也要检测不相邻的词的语义搭配。相邻词可直接通过规则匹配来判断,但词的数量多,不相邻词逐个采用规则匹配的流程复杂,且准确率不高。

[0135] 设i和j分别表示待校文本句中任意两个词。采用互信息算法,对i和j这两个词,通过对比语句语料库相同词语的搭配,用公式计算i和j的互信息值Q(i,j)。用Q(i,j)的大小判断i和j语义的组合情况。计算公式为:$Q(i,j)=\log_2\frac{P(i,j)}{P(i)P(j)}$,P(i,j)为i和j两个词共现的频率,P(i)和P(j)分别为i和j两个词出现的频率。互信息值越大,i和j两个词语义搭配的可能性就越高;反之,i和j两个词语义搭配的可能性就越低。设阈值为0,当Q(i,j)>0时,表示i和j语义组合正确,输出i和j;当Q(i,j)＝0时,表示i和j语义组合不明确,输出i和j,并标注为(?);当Q(i,j)<0时,表示i和j语义组合不正确,输出i和j,并标注为(*)。

[0136] 举例说明,如图4所示。

[0137] 待校文本内容:“打开按扭,就能看电视了”。

[0138] 步骤一:获取待校文本内容:打开按扭,就能看电视了。

[0139] 步骤二:将待校文本切分字符,并分别与固定式语料库、词语语料库、单字词语料库匹配,对待校文本进行分词处理。

[0140] 步骤三:标注词序列,“打开按扭,就能看电视了”标注为1,2,3,…,8。

[0141] 步骤四:将待校文本中的内容与字母语料库、标点语料库、拼音语料库、繁简字语料库比对,判断并显示。

[0142] 步骤五:获取固定式语料库、词语语料库、单字词语料库中位置、词性、语义字段信息,自动标注待校文本中词的位置、词性、语义。

[0143] 步骤六:逐一检测相邻词的组合情况,并用语句语料库比对。通过比对,发现“按”与“扭”在语句语料库中不存在这样的搭配情况。执行步骤七。

[0144] 步骤七:将“按”与“扭”同时进行位置、词性、语义搭配检测,与搭配规则进行比对,发现“按”与“扭”搭配失败,显示“按”和“扭”,并标记为(*)。

[0145] 4.句法结构检测

[0146] 从结构层面分析,汉语语病主要表现为句法成分使用不当、成分残缺或多余、成分搭配不当、语序颠倒、句式杂糅等。

[0147] 4.1建立训练树库语料

[0148] 4.1.1汉语句子分两大类型,主谓句(由主语和谓语构成的句子)和非主谓句(由主谓短语以外的其他短语或词构成的单句),如图5所示。

[0149] 4.1.1.1主谓句语料及示例。

[0150] 表2

[0151]

| 类型 | 主谓句 |
|---|---|
| 名词性主谓句 | 后天星期五。 |

[0152]

| | |
|---|---|
| | 十月一日国庆节。 |
| 动词性主谓句 | 他企图偷越国境。 |
| | 这篇论文发表过。 |
| 形容词性主谓句 | 她的脑子笨得出奇。 |
| | 这房间干干净净的。 |
| 主谓谓语句 | 那本小说我看过五遍。 |
| | 这人什么事都做得出来。 |

[0153] 4.1.1.2非主谓句语料及示例。

[0154] 表3

[0155]

| 类型 | 非主谓句 |
| --- | --- |
| 名词性非主谓句 | 蛇！ |
| | 多美的姑娘！ |
| 动词性非主谓句 | 下雨了。 |
| | 请勿吸烟。 |
| 形容词性非主谓句 | 快！ |
| | 太棒了！ |
| 其它 | 唉！ |
| | 啊！ |

[0156] 4.1.1.3按照主谓句和非主谓句类型，创建树库。

[0157] 4.1.1.3.1新建训练树库，命名为训练树库。

[0158] 4.1.1.3.2增设单句字段(varchar2)、成分字段(varchar2)。

[0159] 4.1.1.3.3录入语料

[0160] 4.1.1.3.3.1主谓句有四种类型，从易到难排列，首先是名词性主谓句(由名词性成分组成，主语与谓语之间通常可加“是”)，结合语句语料库中的语料，从中选取有代表性的500条语料(名词性成分直接作谓语，在语义和句法上有特殊要求：说明日子、天气；在对举的情况下使用，说明职位、身份、学历；说明年龄、数量、容貌、价格、籍贯、所属等)，录入训练树库单句字段中。

[0161] 4.1.1.3.3.2形容词性主谓句由主语与形容词性成分组成，谓语的核心词为形容词，现已从语句语料库中选取了1000条语料(形容词性成分作谓语的情况包括：形容词单独使用；形容词+补语；状语+形容词；状语+形容词+补语；两个及以上形容词并列；形容词+“的”)，录入训练树库单句字段中。

[0162] 4.1.1.3.3.3主谓谓语句由一个大主语和一个由主谓短语充当的谓语构成，大主语与紧跟的主谓短语一般有五种关系：施事||受事+谓语；受事||施事+谓语；大主语与小主语有领属关系；谓语里含有复指大主语的成分；介词/状语+大主语+主谓短语。小谓语可以是名词性成分、也可以是形容词性成分或动词性成分等，现已从语句语料库中选取包含上述类型的1500条语料，录入训练树库单句字段中。

[0163] 4.1.1.3.3.4动词性主谓句分五类(包括述宾结构、述补结构、连谓结构、兼语结构和特殊句式)，现已从语句语料库中选取2500条语料(动词性成分作谓语包括：主语+动词；主语+动词+宾语；主语+动词+动态助词；主语+状语+动词+定语+宾语；主语+状语+动词+补语+宾语；主语+动词+补语；主语+状语+动词+补语+宾语；主语+动词+宾语+宾语；连谓结构充当谓语；兼语结构充当谓语)，录入训练树库单句字段中。

[0164] 4.1.1.3.3.5非主谓句包括四种类型，名词性非主谓句(核心词为名词，包括：事物呈现或突然发现以引起注意；称呼或呼唤某人；时间、地点、环境；数量、价格、籍贯、所属等)从语句语料库中选取了500条语料，录入训练树库单句字段中。

[0165] 4.1.1.3.3.6动词性非主谓句(核心词为动词，包括：单个动词；动词+宾语+助词/

语气词;状语+动词;状语+动词+宾语;状语+动词+定语+宾语;状语+动词+补语+宾语;动词+补语;状语+动词+补语+宾语;动词+宾语+宾语;连谓结构;兼语结构)从语句语料库中选取了1500条语料,录入训练树库单句字段中。

[0166] 4.1.1.3.3.7形容词性非主谓句(核心词为形容词,包括:形容词单独使用;形容词+补语;状语+形容词;状语+形容词+补语;两个及以上形容词并列;形容词+“的”)从语句语料库中选取了1000条语料,录入训练树库单句字段中。

[0167] 4.1.1.3.3.8另外,还包含叹词、拟声词等。录入《现代汉语词典》所标记的常用拟声词和叹词,共67条,录入训练树库单句字段中。

[0168] 4.1.1.3.3.9最终一共构建8567条汉语训练树库。

[0169] 4.1.1.3.4标注成分

[0170] 汉语单句结构复杂,检测单句结构主要检测单句内部各个成分,对此,在构建训练树库之后,需标注句中的句法成分。

[0171] 利用句法分析器(FDG)自动标记训练树库中的句法成分,再人工逐条核对,录入到成分字段中,如:“老李叫小明买东西”。

[0172] 如图6所示,“老李”是句子的主语,“叫小明买东西”是谓语,“叫”是整个句子的核心动词,所以作为根节点,“小明买东西”是主谓句,“小明”是主语,“买”是谓语句中的核心动词,“东西”是宾语。

[0173] 4.2句法结构检测方法,如图7所示。

[0174] 4.2.1待校文本句子分层。获取前期处理过程中的句子分层结果。

[0175] 4.2.2获取前期处理过程中的分词结果(自动分词步骤依次输出的固定式、词语、单字词)及3.2词序列。

[0176] 4.2.3标记待校文本句中虚词,与单字词语料库位置字段比对,自动标注虚词位置。

[0177] 4.2.4提取句法成分。

[0178] 4.2.4.1判断单句成分。

[0179] 在分词基础上,从左往右依次切分出长度不等(从第一个词开始,依次切分出一个词、两个词、三个词…直到Z个词)的字符串,设x为切分出的字符串,设t为x的句子成分,使用概率分布的极大似然法计算在x字符串出现的情况下,t的经验概率:$\tilde{P}$(t|x)=$\frac{freq(x,t)}{\sum_{x,t} freq(x,t)}$,将x字符串与训练树库中成分字段比对,freq(x,t)表示字符串x及对应的句子成分t在训练树库中出现的次数。设阈值为0.2,通过测试,当$\tilde{P}$(t|x)=0.2时,字符串x充当句子中相应成分的可能性极高,当$\tilde{P}$(t|x)>0.2时,就可以判断字符串x是句中的句子成分。如果$\tilde{P}$(t|x)≥0.2,则切分出字符串x,并标记成分名称;如果$\tilde{P}$(t|x)<0.2,表示该字符串x不是句中句子成分,则判断下一个字符串,直到所有成分标记完成,输出未被标记的成分名称,标记为(-*-)。

[0180] 4.2.4.2设成分序列为$X_1$,$X_2$,$X_3$…,$X_n$,将标记的成分与训练树库成分字段比对,自动生成树库。

[0181] 4.2.4.3获取待校树库节点。

[0182] 4.2.5成分搭配检测。遍历扫描树库中$X_1$,$X_2$,$X_3$…,$X_n$成分,并与训练树库成分字段匹配,具体流程如图8所示,包括:

[0183] 步骤一:依据4.2.4.2所生成的树库,查找该树库的根节点。

[0184] 步骤二:访问该节点。

[0185] 步骤三:判断该节点是否有未访问的子节点。如果有,执行步骤四;如果没有,执行步骤五。

[0186] 步骤四:访问最左侧未被访问的子节点,并将该节点与根节点组合搭配,与训练树库成分字段比对。如果正确,则输出该节点对应成分,执行步骤二;如果错误,则输出该节点对应成分并标记为(-*-),执行步骤二。

[0187] 步骤五:判断该节点是否为根结点。如果是,执行步骤六;如果不是,执行步骤七。

[0188] 步骤六:将该节点与训练树库成分字段比对。如果正确,则输出该节点对应成分;如果错误,则输出该节点对应成分并标记为(-*-)。

[0189] 步骤七:返回该节点的父节点。执行步骤三。

[0190] 5.语义表达检测

[0191] 文本中因表达造成的语病主要体现在句中语义搭配上,包括语义搭配不当、歧义、不合逻辑等。

[0192] 5.1汉语是语义型语言,其组合搭配是按照一定的语义规则来进行的。句法成分是有序的,语义成分是无序的,且汉语句法结构与语义关系之间存在复杂的“一对多”“多对一”的对应关系。因此,仅根据句法结构建立训练树库还不够完整,需将二者结合起来。如图9所示,其中左图(句法结构)、右图(语义关系)。两个句子从句法结构分析,都是主谓结构,但语义关系不同。

[0193] 再如图10所示,其中左图(句法结构)、右图(语义关系)。两个句子从语义关系分析,都是事物与性状关系,但句法结构不同。

[0194] 5.2构建语义训练树库

[0195] 5.2.1新建训练树库,命名为语义训练树库。

[0196] 5.2.2增设单句字段(varchar2)、成分字段(varchar2)、规则字段(varchar2)。

[0197] 5.2.3录入语义训练树库语料

[0198] 将4.1.1.3.3所建立的8567条语料,录入语义训练树库的单句字段中。

[0199] 5.2.4标注语义成分

[0200] 利用句法分析器(FDG)自动标记语义训练树库,再人工标记核心动词和语义格(参见表4),录入到成分字段中。例:“老李叫小明买东西”。

[0201] 如图11所示,“老李”是句子的施事,“叫”是句子核心动词(为根节点),“小明”是“叫”的受事,也是“买”的施事,所以既是“叫”的子节点,又是“买”的父节点,“买”是谓语句中的核心动词,“东西”是“买”的受事。

[0202] 5.3语义格特征分析

[0203] 汉语组合搭配是按照一定的语义规则来进行的。语义格组成和搭配也有一定规律可循。

[0204] 5.3.1汉语语义格系统分层

[0205] 表4

[0206]

| 汉语语义格系统 | | |
|---|---|---|
| 第一层 | 第二层 | 第三层 |
| 角色 | 主体 | 施事、当事、领事 |
| | 客体 | 受事、客事、结果 |
| | 邻体 | 与事、同事、基准 |
| | 系体 | 系事、分事、数量 |
| 情景 | 凭借 | 工具、材料、方式 |
| | 环境 | 范围、时间、处所、方向 |
| | 根由 | 依据、原因、目的 |

[0207] 5.3.2语义格基本特征

[0208] 表5

[0209]

| 格名称 | 特征 |
|---|---|
| 施事 | 处于主语位置时，不带格标；处于其他位置时，常带格标（被、让、给、由、归） |
| 当事 | 不带格标，不进入被动句 |
| 领事 | 与客体有领属关系；人或事物的整体 |
| 受事 | 处于宾语位置和受事主语句主语位置时，不带格标；处于其他位置，常带格标（把、将、对） |
| 客事 | 非自发动作行为所涉及的直接客体；不带“把”类格标，不进入“把”字句 |
| 结果 | 可带“把”“将”格标；做宾语时，后面可以加“成”“出”“起来”等 |
| 与事 | 事件中间接客体，被给予者或者被取得者；帮助、服务的对象；常带格标“给”“向”“替”“跟”“为”等 |

| | |
|---|---|
| 同事 | 表现为汉语中的协同动词，常带格标“跟”“和”“同”“连”“除了”；在动作行为动词之后作宾语，可以提前 |
| 基准 | 带格标介词“比” |
| 系事 | 主体的类别、身份、角色；充任系事的动词常为“是”“当”“任”“为”“作为”“成为”“演”“姓”等 |
| 分事 | 领事的组成部分；前常出现领属关系动词“有” |
| 数量 | 数词+量词；对应语料库中“mg”符号的词 |
| 工具 | 动词+工具可变换为“用”/“拿”+工具+动词 |
| 材料 | 句中表材料或物资的词，所带格标“用”“拿”“由”“把” |
| 方式 | 格标“用”“以”；动词+方式宾语可变换为“以”/“用”+方式宾语+动词 |
| 范围 | 所带格标“关于”“就”“在……方面”“在……上”“在……下”“在……中”等 |
| 时间 | 表时间的词，对应语料库中“nt”符号的词 |
| 处所 | 表处所的词，对应语料库中“nl”符号的词；动词+处所词可转换为：介词+处所+动词/动词+介词+处所 |
| 方向 | 所带格标“朝”“向”“往”；表方向的词，对应语料库中“nd”符号的词 |
| 依据 | 所带格标“根据”“据”“按”“按照”“遵照”“照”“依照”“依据”“依”“凭”“靠”“论” |
| 原因 | 所带格标“因为”“因”“为”；和动词组合充当宾语时可变换为：格标+原因+动词 |
| 目的 | 所带格标“为”“为了”；和动词组合充当宾语时可变换为：格标+目的+动词 |

[0210] 语义格基本特征的描述是语义搭配的主要规则，通过分析语义格基本特征，可录入规则字段中，以待语义检测时匹配。

[0211] 5.4语义检测方法，如图12所示。

[0212] 5.4.1待校文本句子分层。获取前期处理过程中的句子分层结果。

[0213] 5.4.2获取前期处理过程中的分词结果(依次输出的固定式、词语、单字词)和3.2词序列。

[0214] 5.4.3提取语义成分。

[0215] 5.4.3.1判断单句语义成分。

[0216] 在分词基础上，从左往右依次切分出长度不等(从第一个词开始，依次切分出一个词、两个词、三个词…直到Z个词)的字符串，设x为切分出的字符串，设t为x的语义成分，使用概率分布的极大似然法计算在x字符串出现的情况下，t的经验概率：$\tilde{P}(t|x) =$

$\frac{freq(x,t)}{\sum_{x,t} freq(x,t)}$，将x字符串与语义训练树库中语义成分字段比对，freq(x,t)表示字符串x及对应的语义成分t在语义训练树库中出现的次数。设阈值为0.2，如果$\tilde{P}$（t|x）≥0.2，则切分出字符串x，并标记成分名称；如果$\tilde{P}$（t|x）<0.2，表示该字符串x不是句中语义成分，则判断下一个字符串，直到所有成分切分并标记完成，输出未被标记的成分名称，标记为(-*-)。

[0217] 5.4.3.2设语义成分序列为$X_1$,$X_2$,$X_3$…,$X_n$，将标记的语义成分与语义训练树库成分字段比对，自动生成树库。

[0218] 5.4.3.3获取待校树库节点。

[0219] 5.4.4语义成分搭配检测。利用语义训练树库及其规则字段，遍历执行$X_1$,$X_2$,$X_3$…,$X_n$的语义搭配检测。具体流程如图13所示，包括：

[0220] 步骤一：依据5.4.3.2待校文本所生成的树库，查找该树库的根节点。

[0221] 步骤二：访问该节点。

[0222] 步骤三：判断该节点是否有未访问的子节点。如果有，执行步骤四；如果没有，执行步骤五。

[0223] 步骤四：访问最左侧未被访问的子节点，并将该节点与根节点组合搭配，与语义训练树库成分字段和规则字段比对。如果正确，输出该节点对应语义成分，执行步骤二；如果错误，则输出该节点对应语义成分并标记为(-*-)，执行步骤二。

[0224] 步骤五：判断该节点是否为根结点。如果是，执行步骤六；如果不是，执行步骤七。

[0225] 步骤六：将该节点与语义训练树库成分字段和规则字段比对。如果正确，则输出该节点对应成分；如果错误，则输出该节点对应成分并标记为(-*-)。

[0226] 步骤七：返回该节点的父节点。执行步骤三。

[0227] 本发明另一方面还提供一种与上述方法步骤一一对应的基于规则和语料库的汉语语病自动检测设备，包括文本获取装置、句子分层装置、正向切分字符串装置、逆向切分字符串装置、自动分词装置和语病检测装置。优选地，语病检测装置包括错别字检测装置、用词不当检测装置、句法结构检测装置、语义表达检测装置，语病检测装置可以包括这四种检测装置中的一种或几种。

[0228] 本领域普通技术人员可以理解，上述实施例的各种方法中的全部或部分步骤是可以通过程序指令相关的硬件来完成的，该程序可以存储于计算机可读存储介质中，存储介质可以包括：只读存储器(ROM,Read Only Memory)、随机存取记忆体(RAM,Random Access Memory)、磁盘或光盘等。

[0229] 本发明并不局限于前述的具体实施方式。本发明扩展到任何在本说明书中披露的新特征或任何新的组合，以及披露的任一新的方法或过程的步骤或任何新的组合。

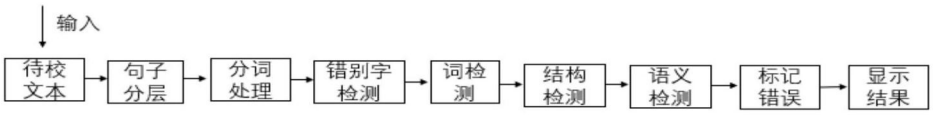

输入
待校文本
句子分层
分词处理
错别字检测
词检测
结构检测
语义检测
标记错误
显示结果

图1

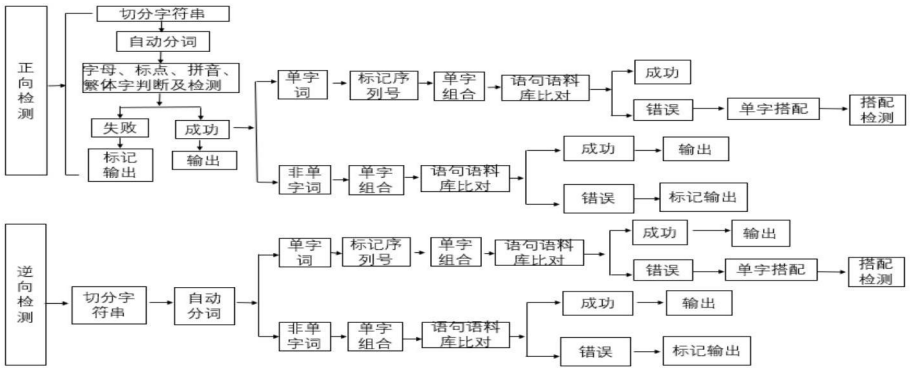

正向检测
切分字符串
自动分词
字母、标点、拼音、繁体字判断及检测
失败
成功
标记输出
输出
单字词
标记序列号
单字组合
语句语料库比对
成功
错误
单字搭配
搭配检测
非单字词
单字组合
语句语料库比对
成功
输出
错误
标记输出
逆向检测
切分字符串
自动分词
单字词
标记序列号
单字组合
语句语料库比对
成功
输出
错误
单字搭配
搭配检测
非单字词
单字组合
语句语料库比对
成功
输出
错误
标记输出

图2

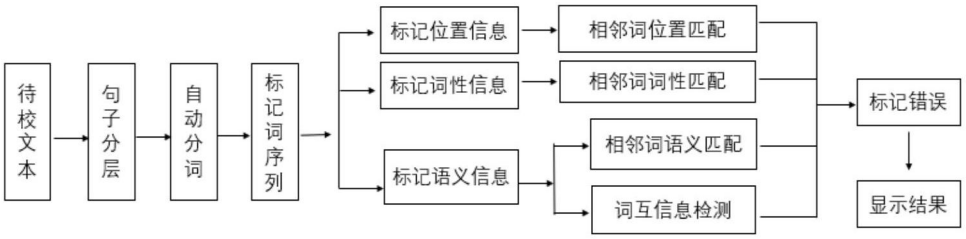

待校文本
句子分层
自动分词
标记词序列
标记位置信息
标记词性信息
标记语义信息
相邻词位置匹配
相邻词词性匹配
相邻词语义匹配
词互信息检测
标记错误
显示结果

图3

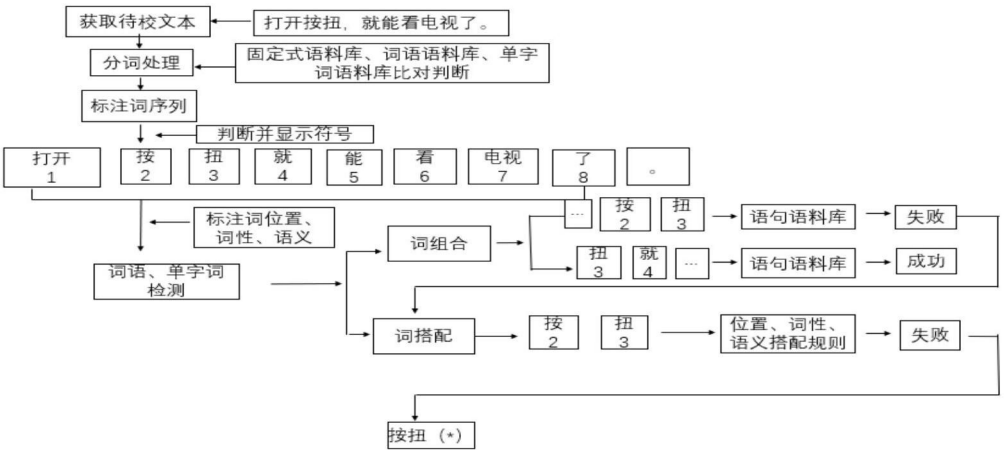

获取待校文本
打开按扭，就能看电视了。
分词处理
固定式语料库、词语语料库、单字词语料库比对判断
标注词序列
判断并显示符号
打开 1
按 2
扭 3
就 4
能 5
看 6
电视 7
了 8
。
标注词位置、词性、语义
词语、单字词检测
词组合
…
按 2
扭 3
语句语料库
失败
扭 3
就 4
…
语句语料库
成功
词搭配
按 2
扭 3
位置、词性、语义搭配规则
失败
按扭（*）

图4

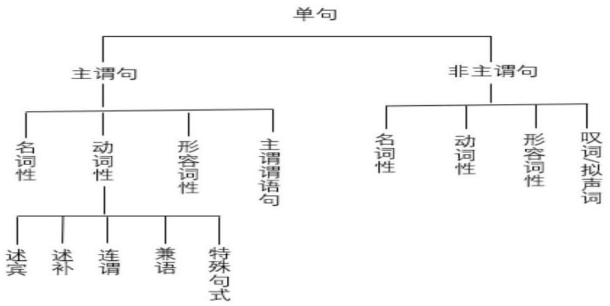

单句
主谓句
非主谓句
名词性
动词性
形容词性
主谓谓语句
名词性
动词性
形容词性
叹词拟声词
述宾
述补
连谓
兼语
特殊句式

图5

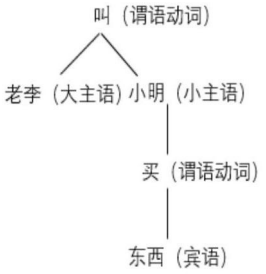

叫（谓语动词）
老李（大主语）
小明（小主语）
买（谓语动词）
东西（宾语）

图6

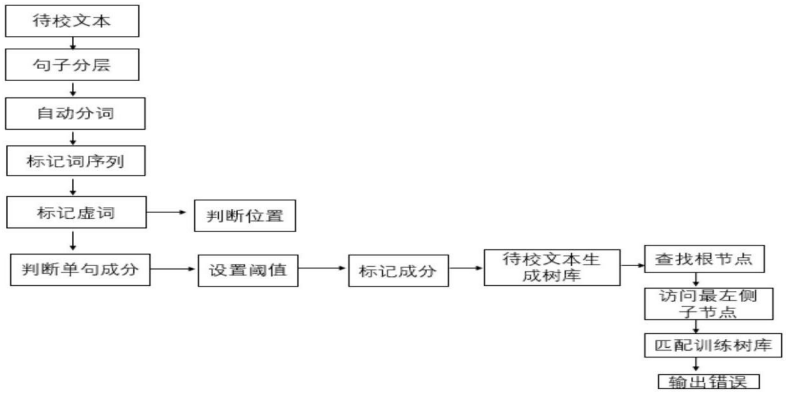

待校文本
句子分层
自动分词
标记词序列
标记虚词
判断位置
判断单句成分
设置阈值
标记成分
待校文本生成树库
查找根节点
访问最左侧子节点
匹配训练树库
输出错误

图7

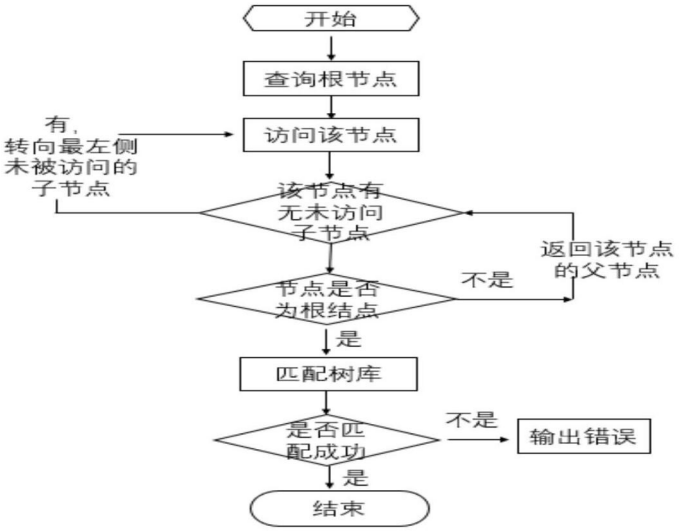

开始
查询根节点
访问该节点
有，
转向最左侧
未被访问的
子节点
该节点有
无未访问
子节点
返回该节点
的父节点
不是
节点是否
为根结点
是
匹配树库
不是
是否匹
配成功
输出错误
是
结束

图8

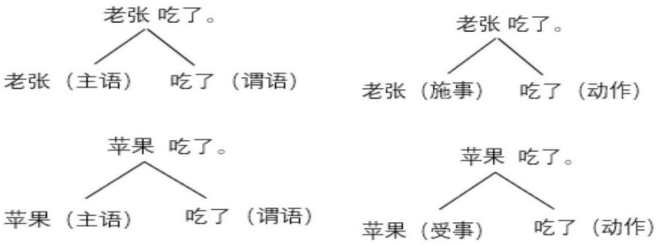

老张 吃了。
老张（主语） 吃了（谓语）
老张 吃了。
老张（施事） 吃了（动作）
苹果 吃了。
苹果（主语） 吃了（谓语）
苹果 吃了。
苹果（受事） 吃了（动作）

图9

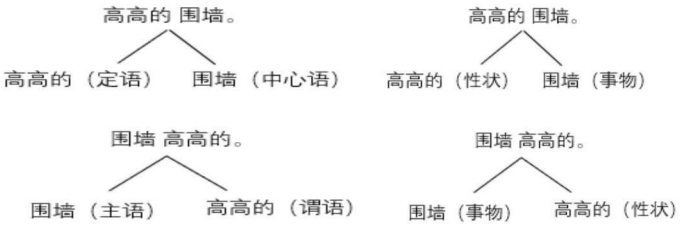

高高的 围墙。
高高的（定语） 围墙（中心语）
高高的 围墙。
高高的（性状） 围墙（事物）
围墙 高高的。
围墙（主语） 高高的（谓语）
围墙 高高的。
围墙（事物） 高高的（性状）

图10

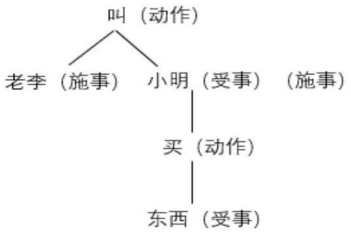

叫（动作）
老李（施事） 小明（受事） （施事）
买（动作）
东西（受事）

图11

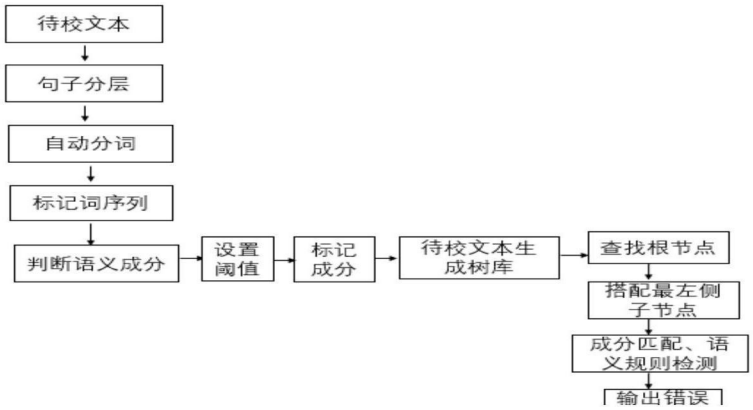

待校文本
句子分层
自动分词
标记词序列
判断语义成分
设置阈值
标记成分
待校文本生成树库
查找根节点
搭配最左侧子节点
成分匹配、语义规则检测
输出错误

图12

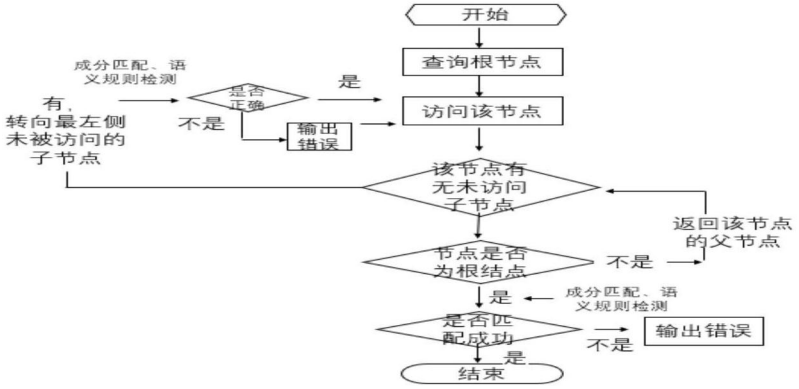

开始
查询根节点
访问该节点
成分匹配、语义规则检测
是否正确
是
不是
输出错误
有，转向最左侧未被访问的子节点
该节点有无未访问子节点
返回该节点的父节点
节点是否为根结点
不是
是
成分匹配、语义规则检测
是否匹配成功
输出错误
不是
是
结束

图13