汉语语音理解中自动纠错系统的研究*

李晶皎 张 琍 姚天顺

(东北大学信息科学与工程学院 沈阳 110006)

摘要 根据汉语语音的特点,通过总结在连续汉语语音识别的汉字序列中出现错误的规律,写出相应的用于查错和校正的词法和句法语义规则.利用"词汇语义驱动"的分析方法,找出汉字序列中的错误并校正,最终得到正确的汉字序列.

关键词 汉语语音理解,词法语义驱动,查错,校正,词法分析,句法语义分析,词法规则,句法语义规则.

中图法分类号 TP391

计算机汉语连续语音识别是要把输入的语音序列转换为正确的文字序列.它通常需要经过语音识别和语音理解的两级转换,语音识别是将输入的语音波形转换成拼音序列,语音理解是将拼音序列转换成文字序列.由于汉语语音的调值具有辨意作用,考虑汉语拼音的声调大约有1 300个拼音,而汉字有近万个,使得语音理解很困难,所给出的汉字序列经常出现错误[1~3].我们的自动纠错系统就是要把语音理解给出的汉字序列中的错误查找出来,并校正输出正确的汉字序列.我们知道,人能听懂别人讲的话句,不全靠听懂每一个字或词的发音,有时并未听清某些字音,却能理解整个话句,在很大程度上是根据谈话的各字音的前后关联以及相应的语法知识来理解整个句子的.因此,计算机自动纠错处理的关键在于综合运用各种可用的声学、语音学、词法、句法、语义以及上下文等多种知识和约束来消除转换中的错误.

在自动纠错系统中,词法、句法和语义等知识是在词典中描述的,约束是通过规则实现的,所用的分析方法是词汇语义驱动[4,5].

1 错误的来源和纠错的种类

连续汉语语音处理中的错误的主要来源有:

(1) 连续汉语语音识别产生的错误.例如:"这使我的房子".这句话正确的应是:"这是我的房子".这是由于语音的调值错误造成的,"使"(shi3)的调值是上声,"是"(shi4)的调值是去声.

(2) 拼音汉字转换产生的错误.例如:"他期在哭".这句话正确的应是:"他妻在哭".这是由于拼音中qi1的同音字太多,不能区分是哪一个qi1产生的.

一个字或词之所以被认为用错了,是因为它与其所在的上下文环境不相适应,即词法、句法和语义关系搭配不当造成的.通过对大量错误现象的研究,以及我们现有的基于词汇语义驱动的语言理解开发平台,我们设计的系统主要是自动纠正下面3种错误.

(1) 词法错误:分词错误或错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.例如:"这里(礼)物的确贵".正确的分词应该是:"这·礼物·的确贵",而不是:"这里·物·的确贵".

(2) 句法错误:某些错误虽然破坏了原词的结构,但却能与其前后字构成词,或者该单字本身就是词,通过

* 本文研究得到国家自然科学基金、国家863高科技项目基金和国家教委博士点基金资助.作者**李晶皎**,女,1964年生,在职博士生,副教授,主要研究领域为语音识别,计算语言学,智能人机接口.**张琍**,女,1961年生,博士生,主要研究领域为计算语言学,机器翻译.**姚天顺**,1934年生,教授,博士生导师,主要研究领域为计算语言学,人工智能,智能人机接口.

本文通讯联系人:李晶皎,沈阳 110006,东北大学信息科学与工程学院

本文1997-12-15收到原稿,1998-04-03收到修改稿

句法分析可以发现这种错误.例如:"那素(树)被暴风给刮倒了".该句为"被"字句,主语是受事,而用介词"被"引进施事,句子中的"素"应为"树",词性应该是名词而不应该是形容词.

(3) 语义搭配错误:主要进行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,动词的施事和受事的搭配关系与限制、句型成分短语内部词与词之间搭配等.例如,"他期在哭"."哭"只能是人在哭,"期"是不对的.

2 纠错系统中的定义

在自动纠错系统中,输入是汉字序列,$CI \in W$,W 是汉语词集合,CI 的长度 m 是输入的汉字个数.通过纠错系统对 CI 进行词法分析,产生输出的词序列 CT,$CT \in W$,$CT-\{CT_1,CT_2,\ldots,CT_n\}$,$CT_j \in W$,$j=1,2,\ldots,n$,$n$ 是对 CI 分词的段数,$n<m$.

每一个词 CT_j 都有一个包括词法、句法和语义的属性集合 A_j,由于 CT_j 对应的同音、近音词个数大于或等于 1,所以,纠错系统的语法分析是搜索一个或多个满足语法规则 R 的汉语词序列,通过语法规则 R 的限制最终得到一个无错误汉语词序列 $\{ct_1,ct_2,\ldots,ct_n\}$,$ct_j \in W$.

定义. 汉语自动纠错系统 S-lsd 是一个四元组 $\{W,A,N,R\}$,其中 W 是包含属性的汉语词集合,A 是词的属性集合,$W \cup A$ 构成终止符集,N 是非终止符集,R 是分词规则和句法语义规则集.在语法规则 R 的约束下,S-lsd将长度为 m 的输入汉字序列 CI 经分词规则处理产生汉语词序列 $CT=\{CT_1,CT_2,\ldots,CT_n\}$,$CT_j \in W$,$n$ 是对 CI 分词的段数,CT 经句法语义规则 R 处理,从 E_j(E_j 同音、近音字集合,$E_j \in W$)中确定唯一一个词,最终产生一个无错误的汉语词序列 $\{ct_1,ct_2,\ldots,ct_n\}$.

也就是说,纠错系统 S-lsd 能把一个可能包含错误的汉字序列转换成无错误的汉字序列.本系统所用的汉语分析方法是"词汇语义驱动"方法.每个词的属性集合都是一个复杂特征集,$A_j=\{M_j,S_j,D_j\}$,M_j 是第 j 个词的主结点信息,S_j 是第 j 个词的静态属性表,在 M_j 和 S_j 中除指针信息外,其余信息均放在词典中,D_j 是第 j 个词的动态表,它们是在分析过程中逐渐形成的.词汇语义驱动即是在设置多种结构的复杂特征集基础上,构造一阶逻辑描述式,通过扩展和合一等运算集、词汇语义规则驱动,完成词法、句法和语义结构的分析.

3 知识库的表示

纠错系统中用到的知识有两类,一类是词的知识,即机器词典的组织与表示;另一类是语法知识,它由一系列规则组成.

3.1 词典的组成

本系统所用的词典共收录了 70 124 个词条,根据我们对词典中词条的统计,一字词条为 7 933 个,二字词条为 46 325 个,三字词条为 8 271 个,四字词条为 6 731 个,其余的为五字词~十二字词.为了加速对词典的查找,系统还为词典建立了两种索引,一种是以拼音码为关键字的一级索引,另一种是以领头字为关键字的一级索引.

词典中每个词条由以下几部分组成:词条 lex,带调的拼音 spl,音节数 syll,词性 ccat,下位词性 subcat,前搭配 qdapei,后搭配 hdapei,语法属性 mor,句法属性 syn,语义特征 semfea,语义分类码 rest.

系统中把汉语词性分成 22 个大类,即名词(n),动词(v),形容词(a),代词(r),数词(m),量词(q),时间词(t),助词(u),介词(p),副词(d),状态词(z)等,每个类又分成若干小类,称为下位词性,本词典共有 71 个小类.

3.2 规则的表示

本系统所用的规则与汉英机器翻译系统中的汉语分析器中的规则不同,其特点是:

(1) 机译系统中用的输入汉字序列是完全正确的,这意味着一个句子必然与一个语义分析结果相对应,而纠错系统的输入汉字序列可能有错误,分析过程正是查找错误的过程.错误的查找是通过总结语音理解给出的汉字序列中经常出现错误的规律,填写词法分析和句法语义分析的规则实现的.

(2) 机译系统中的汉语分析器是通过构造词法结构、句法结构,最终构造出用于生成器的语义结构;而纠错系统最终是确定并校正输入汉字序列的错误,并不试图建立句子的语义结构.这意味着并不要求解决句子中所

有的语法问题,查错和校正是纠错系统的目的.

(3) 机译系统中的汉语分析器必须处理某些歧义结构,例如,"公园里有三个幼儿园的孩子",在句法分析中有两种可能的分析结果,一是指孩子共有 3 个,他们都是幼儿园的孩子;另一种是指孩子来自 3 个幼儿园.但在纠错系统中的句型成分分析只是把"三个幼儿园"分析为"孩子"的定语,并不去深究该短语的内部细微结构.

本系统所用的规则按语法分析可分成两类:一类是词法规则,用 Rw 表示;另一类是句法语义规则,用 Rg 表示.按规则适用范围可分为:共性规则和个性规则.因此,$R=\{Rws,Rwc,Rgs,Rgc\}$,其中 Rws 为个性词法规则,Rwc 为共性词法规则,Rgs 为个性句法语义规则,Rgc 为共性句法语义规则.此外,为了提高规则描述语言的能力,本系统还设计了规则描述语言函数,以@开头,例如:@$SEARCH$().

我们通过对语音理解所得的汉字序列的研究,总结并归纳出汉字序列中错误的规律,写出一系列词法分析规则和句法语义分析规则.在进行规则匹配时,首先查找个性规则,然后查找共性规则.

我们的个性规则是以词条为索引的,它的结构是

Rule={index_word, condition_window, perform_window}.

共性规则是没有索引的,它的结构是

Rule={condition_window, perform_window}.

其中 condition_window={[condition]},perform_window={[performance]},[ ]表示重复必选项.

比如,有这样一句话:"小鸟一直飞了两小时".语音理解给出的句子却是:"小鸟一直飞了量小时".为了检测并校正这类错误,我们可以写出这样一条句法语义的共性规则 Rgc:^ −ccat. m+(ccat. t)=>^. 改. ccat. m and @$SEARCH$(^ spl,ccat. m),其含义是:如果当前词之后的词性是时间词,而当前词的词性不是数词,那么,将当前词的词性改为数词,并且按拼音索引查找与当前词拼音相同且词性是数词的词条,用该词条替换当前词词条.

4　纠错系统的实现

汉语语音理解中的纠错系统的处理主要有词法分析和句法语义分析两种,系统构成如图 1 所示.

4.1　自动分词和词性兼类处理

首先对输入的句子加绝对切分标志,主要是对句中出现的标点等前后加切分标志以加快分词速度,然后取两个绝对切分标志之间的字,用最大匹配法和 2-3-1 法,按切细不切粗的原则进行自动分词.同时要参考语音识别所给出的分段标志.例如,语音识别给出的汉字序列是一个含有错误的句子:"门口是一刻槐树".句子的分词结果是

门口(s,s)是(v,a)一刻(d)槐树(n).(g)

其中 s 为处所词.

词性兼类处理是一个涉及语境的难题,因为汉语中有许多词在不同的语境下可以充当不同的词类,而且含义也不同.为此,我们用规则来解决词性兼类问题.经词性兼类处理后例句为

门口(s,sss)是(v,vv6)一刻(d,dd3)槐树(n,nn1).(g,gg1)

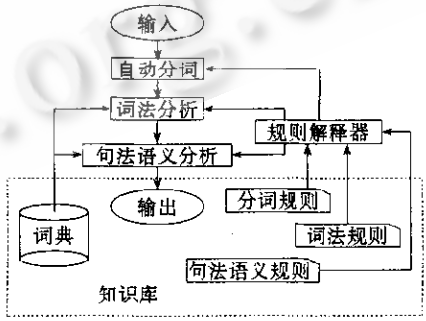

图1　纠错系统结构图

匹配的规则为

Rs:^ ('是')+−gg1=>@$SETMARK$(vv65)

其中@$SETMARK$(C_i)为规则描述语言函数,C_i 为上位词性、下位词性或概念号,将当前词满足工作单为 C_i 的复杂特征集 A_i 中动态表 D_i 的 flag 标志置为 1.

4.2　词法错误的查找与校正

在词法分析中,我们查找词法错误,如果查找到了词法错误,那么立即校正;如果没有错误,则进行下一步的句法语义分析.我们总结语音识别和语音理解中经常出现的词法错误的规律,填写词法分析的共性规则 Rwc 和

个性规则 *Rws*. 错误的查找是通过匹配规则实现的,首先匹配个性规则,然后匹配共性规则. 例如:"这使我的房子". 匹配的个性规则 *Rws* 是

$$\text{subj}+\hat{}\ \text{使}+\text{obj}+(-\text{semrela.comp})=>\hat{}\ .\text{改.lex.是}$$

这条规则的意思是:使动句中没有补语,将"使"改为"是". 校正后的句子为:"这是我的房子".

4.3 句法语义错误的查找与校正

在句法语义分析中,我们检查句子是否有句法错误和语义错误. 如果有错误,根据匹配的规则加以校正. 例如:"他期在哭". 匹配的共性规则 *Rgc* 是

$$\hat{}\ (\text{synrela.subj},(\text{ccat.v};\text{ccat.t}))=>\hat{}\ .\text{改.(rest.1111)},@LOOK\_UP(\hat{}\ )$$

这条规则的意思是:主语是动词或量词,改其语义码为 1111,在词典中查询相应的词条. 校正后的句子为:"他妻在哭".

再比如:"从这个月开始征收薪水". 匹配的个性规则 *Rgs* 是

$$\hat{}\ (\text{ccat.v},\text{lex.征收})+(\text{spl.xinshui})=>:=(\hat{}\ \text{rest},1332),@LOOK\_UP(\hat{}\ 1)$$

这条规则的意思是:"征收"后的宾语的语义码应为 1332,在词典中查询"征收"右侧第 1 个拼音为 xinshui 的相应词条. 校正后的句子为:"从这个月开始征收新税".

5 实验结果

为了有效地对系统进行验证,我们用 141 个语音识别系统所给出的句子作为本系统的输入(其中有 62 个句子有错误),经过纠错系统的处理,查出并校正了 25 个句子,其中词法错误为 15 个,句法语义错误为 10 个. 下面给出部分运行实例.

例 1:"建北平为首度". 匹配的共性规则 *Rwc* 是

$$-@WORD\_EXIST(\hat{}\ )=>@LOOK\_UP(\hat{}\ )$$

这条规则的意思是:句子中的某个词条或短语不成词,重新按拼音查询词条. 校正后的句子为:"建北平为首都".

例 2:"在美国长有犯罪事件发生". 匹配的个性规则 *Rws* 是

$$\hat{}\ (\text{spl.chang},(\text{ccat.a};\text{ccat.v}))+(\text{certer.ccat.v})=>\hat{}\ .\text{改.(lex.常,ccat.d)}$$

这条规则的意思是:遇到"长"修饰中心动词的时候,改为"常". 校正后的句子为:"在美国常有犯罪事件发生".

例 3:"现在的年轻人以留长发为每". 匹配的共性规则 *Rwc* 是

$$\hat{}\ (\text{style.}\text{“以...为”})+(\hat{}\ \text{sentence.comp.}-(\text{ccat.n};\text{ccat.a}))=>\hat{}\ \text{sentence.comp.改.}$$
$$((\text{ccat.n};\text{ccat.a})),@LOOK\_UP(\hat{}\ \text{sentence.comp})$$

这条规则的意思是:"以...为"句式的补语不是名词或形容词,重新按拼音查询词性为名词或形容词的词条. 校正后的句子为:"现在的年轻人以留长发为美".

例 4:"工人们采取罢工以抗拒治本价的残酷剥削". 匹配的共性规则 *Rwc* 是

$$-@WORD\_EXIST(\hat{}\ )=>@LOOK\_UP(\hat{}\ )$$

这条规则的意思是:句子中的某个词条或短语不成词,重新按拼音查询词条. 校正后的句子为:"工人们采取罢工以抗拒资本家的残酷剥削".

例 5:"那素被暴风给刮倒了". 匹配的共性规则 *Rwc* 是

$$\hat{}\ (\text{style.}\text{‘被’})+(\hat{}\ \text{sentence.obj.}-(\text{ccat.n}))=>\hat{}\ \text{sentence.obj.改.(ccat.n)},$$
$$@LOOK\_UP(\hat{}\ \text{sentence.obj})->@NEAR\text{spl}(\hat{}\ )$$

这条规则的意思是:"被"字句中受事的词性不是名词,重新按当前词的拼音查询词性为名词的词条,若找到了则结束,校正成功;否则,触发下一个相近拼音查找函数,根据相近拼音按上述条件重新查询. 本句子中 su 的相近拼音为 shu. 校正后的句子为:"那树被暴风给刮倒了".

例 6:"门口是一刻槐树".匹配的个性规则 *Rgs* 是

^ (ccat.d,lex.刻)+(rest.2223)=>^ ,改.lex.棵

这条规则的意思是:中心词是"树木"类,副词"刻"改为量词"棵".校正后的句子为:"门口是一棵槐树".

本系统目前还不能处理分词错误以及严重的语音识别错误,这将是我们下一步要做的工作.例如,"这里(礼)物的确贵""对这件事要抓(做)好充分准备""这对(堆)老玉米联翩而(连皮)有 30 斤""事情发生在 3.2 时(3 点 20)分".

本纠错系统的特点是:根据语音识别所给出的汉字序列的特点,通过搜集归纳总结错误的种类,填写规则,利用"词汇语义驱动"分析方法,使系统能够自动查找错误并校正,反映出系统具有较高的智能.

参考文献

1 潘凌云,杨长生.拼音、汉字计算机自动转换系统.计算机学报,1990,13(4):271～276
(Pan Ling-yun, Yang Chang-sheng. An auto-system for converting Hanyupinyin to Chinese characters. Chinese Journal of Computers, 1990,13(4):271～276)

2 王晓龙.拼音语句汉字输入系统 InSun.中文信息学报,1993,7(2):45～54
(Wang Xiao-long. Chinese input by Pinyin sentence. Journal of Chinese Information Processing, 1993,7(2):45～54)

3 殷峰,何克抗.语句级拼音-汉字转换系统的设计与实现.计算机研究与发展,1997,34(5):340～345
(Yin Feng, He Ke-kang. Design and implementation of a Pinyin-Chinese character conversion system. Computer Research and Development, 1997,34(5):340～345)

4 Yao Tian-shun. A word-based Chinese language understanding system. International Journal of Pattern Recognition and Artificial Intelligence, 1988,2(1):25～34

5 姚天顺.自然语言理解——一种让机器懂得人类语言的研究.北京:清华大学出版社,1995
(Yao Tian-shun. Natural Language Understanding——A Study of Making a Machine Understand Human Languages. Beijing: Tsinghua University Press, 1995)

Research on Automatic Checking and Confirming Correction for Chinese Speech Understanding

LI Jing-jiao　ZHANG Li　YAO Tian-shun

(*School of Information Science and Engineering　Northeastern University　Shenyang*　110006)

Abstract　According to Chinese speech feature, the authors can sum up the errors that consist in the Chinese characters sequence for continuous Chinese speech recognition, and write the lexical, syntactic and semantic rules for checking and confirming correction. Taking advantage of analysis method of "Lexical Semantic Driven", the authors can check out the errors in Chinese characters sequence, then debug them. At last, correct Chinese characters sequence is obtained.

Key words　Chinese speech understanding, lexical semantic driven, automatic checking, confirming correction, lexical analysis, syntax and semantic analysis, lexical rule, syntactic and semantic rule.