第7卷第11期 软件学报 Vol. 7, No. 11
1996年11月 JOURNAL OF SOFTWARE Nov. 1996

面向屈折语的上下文相关词法分析器*

柯 仑 黄河燕 陈肇雄

(中国科学院计算技术研究所 北京 100080)

摘要 本文提出了一种面向屈折语的上下文相关自然语言词法分析方法,它依据单词的外部形态和其语法、语义特性及上下文成分的特征,建立了形态、语法和语义之间的联系,使得词法分析的结果准确可靠,并且可以利用语言屈折变化的一致性关系,排除不合理的分析结果,减轻语法分析的负担.

关键词 机器翻译,词法分析,自然语言处理.

近年来,计算形态学中广泛采用的方法是使用规则的分析方法,这种方法的思想最初由Koskenniemi提出,被称为"two-level model",后来许多科学家在此基础上作出了不同程度的改进,以适应不同语言变化规律的处理.采用词法规则的分析方法的基本思想是:对不同的自然语言,根据其词形变化规律建立不同的词法规则库,词法分析程序依据规则库中的规则对输入单词进行还原处理,并形成相应的词法特征信息,以供后续处理机制使用.[1]采用规则的词法分析方法,把语言中的形态变化规律作为系统数据进行处理,使得系统具有良好的可维护性和可修改性.然而,语言规律千变万化,不少变化规律是同单词的语法甚至语义特征密切相关的.为此,我们提出了一种规则体系,它建立了单词形态变化与其语法、语义之间的联系,同时还可以根据具体语言特点,使单词与其所处上下文环境发生一定联系,并据此排除不合理的分析结果,提高分析准确率.

1 上下文相关词法规则表示

词法规则有词法分析规则和词法相关规则,下面分别叙述它们的形式.

1.1 词法分析规则

词法规则分规则变形和不规则变形2类.

1.1.1 规则变形规则的一般表示形式为:

〈词缀模式〉→〈适用条件〉,〈还原操作〉,〈属性检查〉,〈词法特征〉

其中〈词缀模式〉描述输入单词的词缀特征,它可以用来表示单词的前缀、中缀和后缀.〈适用

---

* 作者柯仑,女,1970年生,硕士,主要研究领域为机器翻译.黄河燕,女,1963年生,博士,研究员,主要研究领域为语言信息处理,面向对象程序设计,人工智能.陈肇雄,1961年生,博士,研究员,博士导师,主要研究领域为机器翻译,人工智能.

本文通讯联系人:柯仑,北京 100080,中国科学院计算技术研究所

本文1996-06-20收到修改稿

条件〉指输入单词除词缀外的外部形态特征,主要是指单词中特定位置的字符特征,它以〈存在条件〉的形式表示.〈存在条件〉的形式为 $Exist(x,Zone(Dir,(a,b)))$,表示如果在词缀的 Dir 方向的(a,b)区域内存在 x,则条件为真.例如:$Exist(FU,Zone(L,(1,1)))$表示如果词缀左边的第1个字母是辅音时条件为真.〈还原操作〉主要是字符串的替换工作,为了适应不同情况的处理需要,设置了2种替换操作:〈无条件替换〉和〈条件替换〉.

①无条件替换

形式为 $C(\Phi1,\Phi2)$,表示用 $\Phi2$ 串替换单词中的 $\Phi1$ 串,如果 $\Phi1$ 就是〈词缀模式〉中的词缀,可简写为"—".例如,$C(-,\Phi2)$表示以 $\Phi2$ 串替换单词词缀.

②条件替换

形式为 $CC(〈测试区域〉,〈替换表〉)$.其中〈测试区域〉的形式为 $Zone(Dir,(a,b))$,各项含义与上述〈适用条件〉中相同;〈替换表〉的形式为:$[a1/b1|c1|d1...,a2/b2|c2|d2...,.....]$,$ai,bi,ci,di$ 为字符串,"|"表示或关系.即表示在给定〈测试区域〉内,如果存在 ai 串则用 bi 或 ci 或 di 串来替换它.

〈条件替换〉的引入可以用来处理"不连续形变"问题[2],例如德语中的"变音"可以通过设置如下替换解决:$CC(Zone(L,(1,4)),[\ddot{a}/a,\ddot{o}/o,\ddot{u}/u,i/e])$,表示如果词缀左边第1个到第4个字符的区域内,如果有 $\ddot{a}$,则以 a 替换之;有 $\ddot{o}$,则以 o 替换之;有 $\ddot{u}$,则以 u 替换之;有 i,则以 e 替换之.

〈属性检查〉中的属性是指应用该规则还原后的原形单词所应具有的语法、语义属性,语法属性主要是词性,语义属性可以是词的应用领域特性等.

例如:在俄语词法分析规则库中有如下规则:

—в $Exist(FU,Zone(L,(1,1)))$ $C(-,ва)$ $\underline{N\&(HU|AN)}$ $CAS2,CAS4,PL$

其中划线部分$\underline{N\&(HU|AN)}$表示该规则描述的是表示人或动物的名词的2格及4格变化.

整个规则的含义是:如果输入单词以—в结尾,在词尾в左边的第1个字母如果是辅音的话,那么将в替换成ва之后,即得到该单词的原形,这个原形词应是一个名词,并且是人或动物的名称,输入单词是原形单词的复数第2格及第4格形式.

〈词法特征〉是对应于每一种词形变化的特征信息,如名词的格、数的变化,动词的时态变化,形容词、副词的升级变化等.词法分析的特征信息是词法分析的结果,它为后续的语法分析提供了依据.

1.1.2 不规则变形词法规则[3]的表示形式为:〈不规则形〉→〈原形〉,〈词类〉,〈词法特征〉

例如:①went→go,VP,PAST;②best→good,AP,SUPPER

①表示英语中went是动词go的过去式形式;②表示英语中best是形容词good的最高级形式.

1.2 词法相关规则的形式

词法相关规则的基本形式为:〈模式〉→〈操作〉

其中〈模式〉为要处理的局部上下文形式.〈操作〉为对符合〈模式〉的局部语段施以的处理过程,由过程名和参数组成.

例如:$AP(Exist(NP,Zone(R,(1,1))))\rightarrow Consis(GEND,NUM,CASE)$表示如果形容词($AP$)后面紧跟着一个名词($NP$),则形容词和名词应在性($GEND$)、数($NUM$)、格

(*CASE*)上保持一致.

词法相关规则的设置,主要是为了利用语言形变的相关规律,在语法分析之前排除词法分析中的歧义,以减轻语法分析的负担.

2　上下文相关词法分析算法及实现

2.1　词法分析规则的组织和运用

2.1.1　词法规则库的组织

对于规则变化词法规则,把所有含后缀的规则(包括单纯后缀规则及后缀与前缀、中缀的组合规则),按末字符顺序组织,并建立末字符索引表;把所有含前缀的规则(包括单纯前缀规则及前缀与中缀的组合规则),按首字符顺序组织,并建立首字符索引表;如果存在单纯的中缀规则,则把它们放在整个规则库的最前面. 整个规则变化库的组织示意图如图 1.

对于不规则变化词法规则,按照〈不规则形〉的字母顺序排序,以便检索时采用二分法进行查找,提高检索速度.

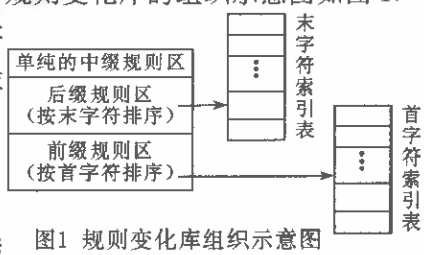

图1　规则变化库组织示意图

2.1.2　词法规则的运用

每一条规则变化词法规则的运用过程如下:

(1)输入单词与〈词缀模式〉相匹配,如果成功,继续执行(2),否则转(6);

(2)检查输入单词是否符合〈适用条件〉,如果符合,继续执行(3),否则转(6);

(3)对输入单词执行〈还原操作〉,得到原形单词;

(4)在字典中检索所得原形单词,如果失败,转(6),否则检查该单词是否满足〈属性检查〉要求,如不满足,转(6),否则执行(5);

(5)将所得原形单词及其〈词法特征〉填入词法分析表;

(6)结束.

应用词法规则对输入单词进行分析的过程如下:

(1)在不规则变化库中,以二分法查找是否存在与输入单词相匹配的规则,如果存在,则将该单词〈原形〉及其〈词类〉和〈词法特征〉填入词法分析表;

(2)将规则变化库中中缀区的所有规则逐一应用到输入单词上;

(3)取输入单词的末字符 T,在末字符索引表中取出以 T 结尾的后缀规则的起始号和终止号,将起始号与终止号之间的所有规则逐一应用到输入单词上;

(4)取输入单词的首字符 H,在首字符索引表中取出以 H 开头的前缀规则的起始号和终止号,将起始号与终止号之间的所有规则逐一应用到输入单词上;

(5)结束.

2.2　词法分析实现算法

对于句子中的每一输入单词,既可能是原形,也可能是变化形,因此词法分析过程首先检索字典,若在字典中有该输入单词,则把该单词和空的词法特征表加入词法分析表中,然后再进入词法规则的处理过程,词法分析的流程图如图 2 所示.

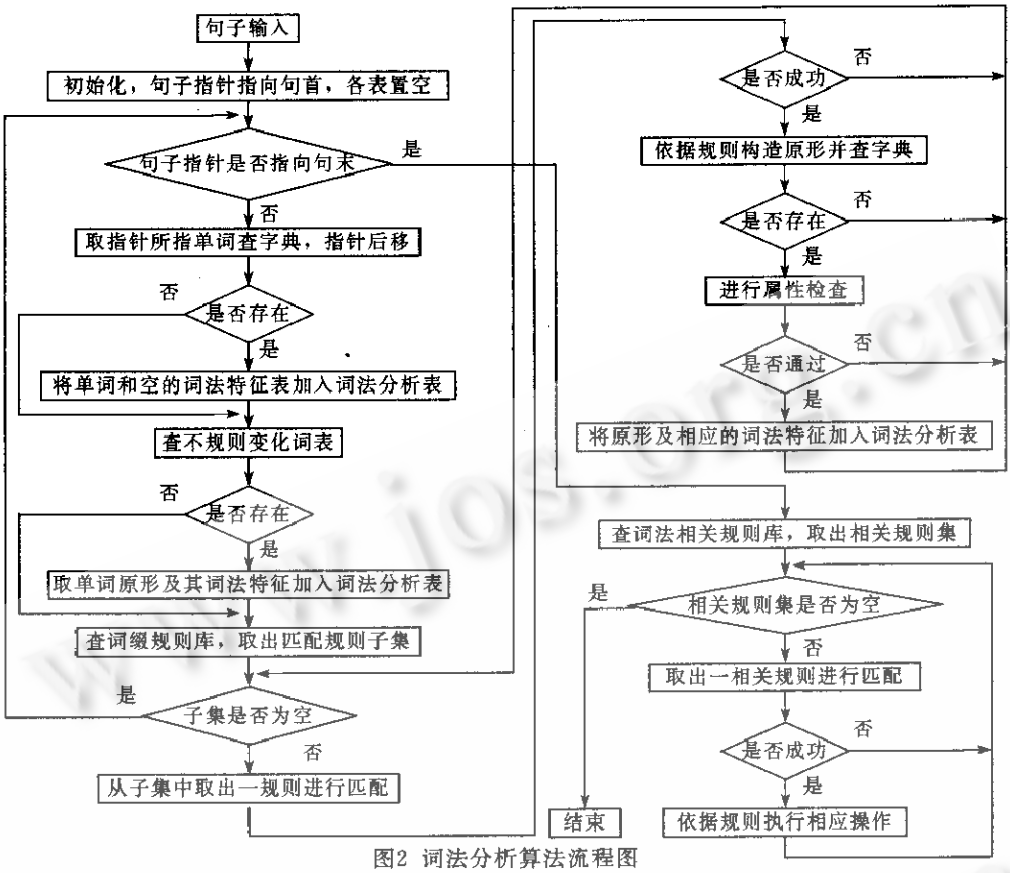

图2 词法分析算法流程图

3 分析结果举例

应用该词法分析方法对屈折性强的语言进行分析，效果良好，如俄语、德语等.

3.1 俄语单词分析实例

3.1.1 包含"语音交替"现象的例子

例1：输入单词 ПИШУ，对应原形单词 ПИСАТЬ(译文：写[未完成体])，获得词法特征(*SINP*1,*VF*).

所用规则：—У →*Exist*({Ч,Ж,Ш,Щ},*Zone*(*L*,(1,1))) *CC*(*L*,(1,1),[Ч/К|Т, Ж/З|Г,Ш/С|Х,Щ/СК|Т]), *C*(—,АТЬ) *VP*&($*NFS*) *SINP*1,*VF*

例2：输入单词 Д ЕВОЧЕК，对应原形单词 Д ЕВОЧКА(译文：小姑娘)，获得词法特征(*CAS*2,*CAS*4,*PLUR*).

所用规则：—ЕК→*Exist*(*FU*,*Zone*(*L*,(1,1))) *C*(—,КА)

NP&(*HU*|*AN*)&$*NAP*&$*NNCAS* *CAS*2,*CAS*4,*PLUR*

需要说明的是，词法规则库中还有另一条规则：

—ЕК→*Exist*(*FU*,*Zone*(*L*,(1,1))) *C*(—,КА)

NP&$*HU*&$*AN*&$*NSIG*&$*NAP*&$*NNCAS* *CAS*1,*PLUR*

输入单词 Д ЕВОЧЕК 之所以不会走这条规则并获得(*CAS*1,*PLUR*)的词法特性，是因

为原形单词 Д ЕВОЧКА 带有 *HU* 属性,不满足〈属性检查〉中 $*HU*(非人)的要求.

3.1.2　需要进行二次变化的例子

例 1:输入单词 НАПИСАН,对应原形单词 НАПИСАТЬ(译文:写[完成体]),获得词法特征(*CAS*6,*SIN*).

首先经一次变化词法规则:

$-\varphi\rightarrow Exist(\{Н\},Zone(L,(1,1)))$　$C(-,НЫЙ)$　$AP\&\$AREL$　$APS,MALE$

处理,得到 НАПИСАННЫЙ(形动词长尾),然后经二次变化词法规则:

$-НННЫЙ\rightarrow Exist(\{А,Я\},Zone(L,(1,1)))$ $C(-,ТЬ)$　$VP\&VACC$　AVB

处理,得到原形单词 НАПИСАТЬ.

3.1.3　利用形变组合一致性的例子

输入词组:КРАСИВОЙ Д ЕВОЧКЕ(译文:美丽的女孩).

对单词 КРАСИВОЙ 单独分析,得到原形词 КРАСИВЫЙ(译文:美丽,好看),词法特征(*FEMA*,*SIN*,*CAS*2,*CAS*5,*CAS*6).

对单词 Д ЕВОЧКЕ 单独分析,得到原形词 Д ЕВОЧКА(译文:女孩),词法特征(*FEMA*,*SIN*,*CAS*3,*CAS*6).

由词法相关规则 $AP(Exist(NP,Zone(R,(1,1))))\rightarrow Consis(GEND,NOM,CASE)$ 处理后,上述 2 个单词的词法特征变为:КРАСИВОЙ——(*FEMA*, *SIN*, *CAS*6);Д ЕВОЧКЕ——(*FEMA*,*SIN*,*CAS*6).

这样,不满足性、数、格组合一致性要求的特征就被排除了.

3.2　**德语单词分析实例**

3.2.1　包含"变音"现象的例子

例 1:输入单词 *Gründe*,对应原形单词 *Grund*(译文:原因),获得词法特征(*CAS*1,*CAS*2,*CAS*4,*PLUR*).

所用规则:$-e\rightarrow Exist(\{\ddot{a},\ddot{o},\ddot{u}\},Zone(L,(1,1)))$ $CC(L,(1,4),\{\ddot{a}/a,\ (\ddot{o}/o,\ (\ddot{u}/u\})$,

$C(-,\varphi)$　　$NP\&\$NEAT\&\$NUGEN$　　$CAS1,CAS2,CAS4,PLUR$

例 2:输入单词 *kälter*,对应原形单词 *kalt*(译文:冷的),获得词法特征(*COM*).

所用规则:$-er\rightarrow Exist(\{\ddot{a},\ddot{o},\ddot{u}\},Zone(L,(1,1)))$ $CC(L,(1,4),\{\ddot{a}/a,\ddot{o}/o,\ddot{u}/u\})$,

$C(-,\varphi)$　　$AP\ COM$

3.2.2　利用形变组合一致性的例子

输入句子:*Die Wirklichkeiten sahen ganz anders aus.*(事实看起来完全不同)

Die:原形单词,查字典,得到译文"这".

Wirklichkeiten:经后缀规则 $-en\rightarrow\varphi C(-,\varphi)$ $NP\ PLUR,CAS1,CAS2,CAS3,CAS4$ 处理后,得到原形单词 *Wirklichkeit*,查字典,得到译文"事实,现实".

Die Wirklichkeiten 经词法分析后,各自所得的属性为:

Die:*FEMA*,*SIN*,*CAS*1,*CAS*4 及 *PLUR*,*CAS*1,*CAS*4.

Wirklichkeiten:*PLUR*,*CAS*1,*CAS*2,*CAS*3,*CAS*4.

它们经过词法相关规则 $ART(Exist(NP,Zone(R,(1,1))))\rightarrow Consis(GEND,NUM,CASE)$(意为冠词与其后名词在性、数、格上保持一致),处理后,各自的属性减少为:

Die:*PLUR*,*CAS*1,*CAS*4. *Wirklichkeiten*:*PLUR*,*CAS*1,*CAS*4. *sahen*:不规则变化动词,经查不规则词表,得到原形词 *sehen*. *ganz*:原形单词,查字典,得到译文"完全". *anders*:原形单词,查字典,得到译文"不同". *aus*:原形单词,又是可分动词前缀,在翻译中与 *sehen* 结合,意为"看起来".

4 结 语

以上介绍了一种结合语法、语义的上下文相关词法分析方法,对语言的词法变化规律给出了较以往更全面、更准确的描述;同时提出了建立词法分析相关规则的方法,使得机译系统能够根据不同语言的特点,利用语言形态变化的相关性,对句子局部进行简单的分析,使后续的语法分析能够把注意力集中在句子整体分析上,以免陷入细节问题难于自拔.

从上面对词法规则形式的介绍可以看出:第1,规则之间是相互独立的,一条规则的修改与增删不会影响其它规则的处理功能;第2,由于规则中设置了〈适用条件〉,使得每条规则的适用范围从外部形态上受到了限制,减少了错误发生的可能性;第3,规则中〈属性检查〉的设置,使经过还原的单词再次经受合法性检查,从语法、语义的角度排除不正确的还原. 规则形式的上述特点,使得词法分析库有良好的可修改性与可扩充性,实现了数据与算法的分离,此外,规则表示形式与具体自然语言无关,因而是通用的. 在机器翻译系统完善过程中,语言学工作者与计算程序设计者的相互配合更容易,当语言学工作者发现例外的语言现象时,只需修改规则中的某些条件,而不必涉及程序的变动.

该词法分析机制已经实现,运用在IMT/EC机译系统的俄—汉及德—汉翻译中,获得了良好的效果.

参考文献

1 陈肇雄. IMT/EC系统总体设计[博士论文]. 北京:中国科学院计算技术研究所,1988.

2 Harald Trost. The application of two—level morphology to non—concatentive German morphology. In: COLING—90, Proceedings of Thirteenth International Conference on Computational Linguistics, Helsinki Finland, 1990. 371~376.

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):93~99.

AN INFLECTED LANGUAGE ORIENTED CONTEXT—SENSITIVE MORPHOLOGICAL ANALYZER

Ke Lun Huang Heyan Chen Zhaoxiong

(*Institute of Computing Technology The Chinese Academy of Sciences Beijing* 100080)

Abstract This paper presents an inflected language oriented context—sensitive morphological analysis approach. In this approach, not only the appearances of words but also their syntactic and semantic attributes are distinguished. Further more, the context in which a word appears is also considered. Thus a more reliable result will be attained. To relax the burden of syntactical analysis, inflection relations between connected words are used to remove unreasonable results.

Key words Machine translation, morphological analysis, natural language processing.