词典规则设计中的分段和继承思想

傅奇芳　聂昱　王斌　戴大为

（武汉大学计算机科学系　430072）

目前大多数机译系统都根据规则库中的规则来刻画语言现象.但是,由于机译规则系 统一般地只注意利用一组元语言来形式化地描述自然语言的普遍现象,而语言中的由约定 俗成或两种自然语言的对比中形成的个性现象,虽然也可以用规则来表示,但是由于不具 有普遍性而被排除在规则库之外.这是影响机器翻译质量的关键所在.

在设计一个用于汉英机器翻译系统的中文词典时,我们的考虑是:我们把规则分为词 典规则和句法规则两大类.词典规则存放在词典中,与具体的词条的某个具体意义相对应 ,句法规则存放在规则库中.句法规则对所有的语言现象起作用,而词典规则只与它相联 系的单词起作用.在设计上采用分段的思想保持词典规则的有序性和一致性, 适当细化规 则,提高词典的通用性和准确性;并引人继承的方法,减少系统开销,提高效率.

通常的汉语语法只规定了汉语的可结合性,它只反映词、短语和句子之间的可搭配或 可修饰关系.传统汉语语法的这种不严密性和过强的生成能力将导致机器翻译中大量的歧 义问题.我们可采用规则描述汉语的可结合过程.这种结合过程直观地可看作由规则组成 的集合.但为分化冗余歧义和微弱歧义.我们将规则集合分割为若干有优先关系的段,并 对每个段的句法规则施加一些限制.为反映结合力的作用,我们定义的句法规则既规定可 结合性,又规定母范畴与子范畴之间的信息传递,还规定子范畴间的结构调整.

规则集合分段须遵从三个原则:(一)由简单到复杂,(二)保持真歧义,(三)消除假歧义。

语法分段固然能大大提高汉语分析和英语生成的效率及准确度,但规则信息的丰富 和细化却无可避免地增加了系统开销和算法复杂度,分段思想的最后实现仍缺乏实践基 础.于是我们又进一步提出继承的思想.所谓继承,就是将若干具备某些相同特征的语 法范畴抽象为一个具有代表性和可重复引用的类, 给类命名和必要的特征描述,然后用 指向类的链来代替原来的语法范畴.我们还进一步将继承发展为缺省继承的思想.

在规则库无法包容自然语言的个性现象时,我们只能考虑从扩大词典的信息容量出 发,由词典提供更加丰富的信息,驱动句法分析器进行工作。受规则形式的启发,我们 引入词典规则,用它来描述单词的特殊用法。

词典规则与句法规则具有相同的形式。一个完整的规则形式如下:

C1C2…Cn→{[<C0i,Condi,Infoi>]m1,head,passage }

其中,m、i、n 为正整数,Ci,C0j 均为语言范畴,Condi 是条件式的集合,Infoi 是信息传 递的集合,head 表示新范畴的中心词在原范畴串序列中的序号,passage 表示该规则可与 哪一个语法段同时使用。规则的含义,直观地讲,对于某个正整数 k,0≤k≤m,如果 Con dk 成立,则语言范畴串 C1C2…Cn 在第 passage 个语法段将构成 C0k,C0k 的中心词是范畴串 C1C2… Cn 中序号为 head 的范畴,C0k 与 C1C2…Cn 之间的信息如何传递将由 Infok 来决定。

本文是以我们在汉英机器翻译方面的研究工作为基础,结合国内外的一些研究成果 ,探讨了词典设计与自然语言及其语法之间的关系,提出了基于分段和继承思想上的词 典规则的概念,使词典规则与句法规则相分离.这样做的好处在于:(1)它利用语法分段 保持了词典规则的有序和一致;(2)它利用语义分类提高了词典规则的实用和准确.