朝鲜语句子分析模型设计要点

社科院民族所　曹雨生

句子分析模型必须根据被处理语言的诸多特点(语音、词法、语法等)进行设计。本分析模型主要由三个关键部分组成:1、词语处理;2、句子分析的句法规则和算法;3、语义分析。本文从语言学角度而非纯处理算法和技术角度出发,介绍了前两部分的语言学理论依据和主要模块功能及其实现。具体论述了三个问题:1、朝鲜语句子结构特点(以书面语为主);2、词语处理;3、句法规则和算法。

一、朝鲜语的主要特点

朝鲜语是粘着型语言,它不同于孤立型的汉语。主要特点:

1、组成句子的词(词组)的类、性、态、形、体、音、义等均服从一定的规律(粘着规律),按规律组词(词组)成句。2、句子结构层次清晰,弄清结构层次及其相互关系,句法规则就较易制定。3、由于粘着型语言的特点,在词的切分上较之汉语要简单得多。句子歧义现象也不像汉语那么复杂。处理就较易。4、句子的基本语序是:主语——补语——谓语结构。句子语法成分主要由词的粘附成分表征。谓词终结形粘附成分结束句子。5、复句是通过谓词连谓形粘附成分的连接构成的,其分析法基于单句分析。

二、词语处理

词语处理的输出供句法分析部分的输入用。其设计主要从两方面着手:构词特点和语法功能。我们在较系统地进行句类统计的基础上,从朝鲜语的五种句子成分入手,借鉴了语音识别中采用的一种"黑板"结构思想,引进到我们的词语处理中,效果较好。词语处理得到的信息主要包括三部分:语法信息、语义信息和标志信息。语法信息主要包括:词语结构信息、形态信息、语法功能信息。语义信息主要包括体词、谓词、修饰词、叹词等的意义特征。标志信息主要特征:词语位置及其变换特征、组词特征等信息。根据上述三部分信息特征,我们把词语处理模型的框架结构设计成一组相应的二维列表。这就是"黑板"(Blackboard)结构的基本思想。该结构是一个开放性、扩展性的模型。修改灵活方便。朝鲜语的粘附成分是封闭的,总共不超过500个,所以模型并不复杂。

三、语法理论及其句法规则

朝鲜语句子分析处理的最大特点在于:句子成分之间无论是表层形式结构关系还是深层语义结构关系,主要决定于词的粘附成分,根据这一特点,在句类统计的基础上,我们总结出了一套语法结构,根据它可制定出相应的句法规则库。朝鲜语的基本语序是:定语+主语+补语+状语+谓语。前两部分组成主语部分,后三部分组成谓语部分。整个单句由两大部分组成。经过句类统计,可归纳出20个表达式。每一个表达式由一个或多个特征信息元组成(词语及粘附成分)。由这一组表达式可建立相应的句法规则库。

四、句子分析算法

句子分析算法的方式通常有两大类:自顶向下(句→分句/子句→词(短语))和自底向上两类。结构也可分为两种:回溯处理和无回溯处理。经过典型例句试验,就朝鲜语而言,虽然这两种方式和结构各不相同,但基于粘着语的特点,其效果大致一致。这和形态不发达的汉语大不相同。我们的系统采用的是自底向上并行算法。这种算法对粘着语效果都较好。另外朝鲜语有一定量的无主句,根据粘着型特点,我们把这一句型归结为所谓"另"主语结构来处理。这样就可把它与有主句一视仁地处理。还有,朝鲜语的语气因由相应的粘附成分表征,亦即有相应的语调和形态标志。因此处理起来就较简单。亦有规律可循。