谈“是……的”句的计算机理解（摘要）

北京语言学院　张宝林

采用检索关键词的方法使计算机在一个文本中提取“是……的”句，会提出大量似是而非的句子。其主要原因是：1、“是……的”句与以“的”字短语作宾语的“是”字句形式相同，都带有“是……的”框架；2、“是……的”句的宾语既可以在“的”字之前，也可以在其后；“是”字句的宾语既可以由中心语不出现的“的”字短语充当，也可由中心语出现的偏正短语充当；3、“是……的”句内部存在“是……的”句（一）和“是……的”句（二）两种情况。由于原因1和2，会使计算机在提取“是……的”句时混入大量的“是”字句；由于原因3，计算机难以把两种“是……的”句区分开来，从而造成了计算机理解此类句子的极大困难。

解决的办法有两种。1、从形式上分解“是……的”句，主要是根据“是……的”框架中间成分的性质分解“是……的”句。在已经逐词标注词性的语料库文本中，①当“是……的”中间是单个的名词、代词、动词、形容词时，句子是“是”字句；②当“是……的”中间是名词性偏正短语时，也是“是”字句；③当“是……的”中间是形容词性偏正短语时，句子是“是……的”句（二）；④当“是……的”中间是动词性偏正短语、连动短语、主谓短语时，句子是“是……的”句（一）。根据上述规则，计算机可以将三类句子区别开来。存在的问题是计算机不能判断各种短语的类型，这就需要我们给出一些构成短语的组合规则，例如：(1) 名词+名词→名词短语；(2) 代词+名词→名词短语；(3) 副词+形容词→形容词短语；(4) 时间名词+动词→动词短语；(5) 处所名词+动词→动词短语；(6) 动词+动词→连动短语；(7) 名词+动词→主谓短语；(8) 代词+动词→主谓短语。有了这些规则，计算机就能对短语的类型加以判断，进而提取出真正的“是……的”句。

运用这种方法，计算机还可以把偏正短语做宾语的“是”字句与“的”在宾语之前的“是……的”句（一）区别开。即在“是……的+名词”这种格式中，当“是……的”中间是名词，代词、形容词、及与之相应的名词短语、形容词短语，以及主谓短语和动词短语时，句子是“是”字句；当“是……的”中间是动词及动词性偏正短语和连动短语时，句子是“是……的”句（一）。

2、从意义上分解“是……的”句。有这样一种歧义现象，即有些带有“是……的”框架的句子既可以看作“是”字句，又可以视为“是……的”句。怎样把这种有歧义的句子与无歧义的句子区别开呢？我们认为可以根据动词的再分类加以区别，即含有“制成义”或“出现义”的动词才可以构成歧义句；而具有“毁损义”或“消失义”的动词不能构成歧义句，只能构成“是……的”句。根据歧义句与非歧义句动词意义上的这种特点，我们可以在计算机中把动词分别定义为“毁损消失义动词”和“制成出现义动词”，有前一类动词的句子是无歧义的“是……的”句，有后一类动词的句子是歧义句。计算机可以据此自动进行分类。