

# 从“是 p, 还是 q”有标选择复句看 合用型关系词的自动识别

尹蔚, 罗进军

(华中师范大学语言与语言教育研究中心, 湖北武汉, 430079)

**摘要:** 自然语言理解是一项复杂而伟大的工程, 其内容的深度及广度是难于想象的。自动识别是其中的一个小小的分支, 这里着重探讨有标复句领域的合用型关系词的自动识别问题。合用型关系词的自动识别是有标复句信息化工程的基础环节之一, 为了有效解决问题, 选取“是……还是……”这对合用型关系词作为研究对象, 以验证研究方法的可行性。

**关键词:** 合用型关系词; 自动识别; 语序制约律; 相互制约律; 句法空间制约律; 亲密度制约律

**中图分类号:** H14

**文献标识码:** A

**文章编号:** 1672-3104(2007)06-0740-04

丁力从本体研究的角度对“是 p, 还是 q”这对合用型关系词进行了详尽的考察<sup>[1]</sup>, 本文则尝试着从应用研究的角度对其进行探索, 具体说来就是从计算机对有标复句关系词的自动识别这个方面予以考察。有标复句关系词的自动识别分为两类, 一类是单用型关系词的自动识别, 另一类是合用型关系词的自动识别, 其中合用型关系词的自动识别是重点, 也是难点。有鉴于此, 我们对“是 p, 还是 q”有标选择复句<sup>①</sup>进行了深入考察, 期待考察结果以及从中得出的规律能够成为其他合用型关系词自动识别的突破口。根据我们的观察, 有四条规律制约着有标复句合用型关系词的自动识别, 它们分别是: ① 语序制约律; ② 相互制约律; ③ 句法空间制约律; ④ 亲密度制约律。下面我们分别从这四个方予以解释。

## 一、语序制约律

合用型关系词在语表形式上存在着一种相对稳定的先后顺序, 比如“是 p, 还是 q”有标选择复句中, “是”相对居前, “还是”相对居后, “是”跟“还是”一般情况下比较容易构成合用型关系词。但是, 如果“是”

居于“还是”之后, 那“是”跟“还是”一般不会构成合用型关系词。比如:

(1) “不必太过‘世故’”, 这无论是对青年人, 还是对老年人, 都是很有教益的。(《人民日报》1983年08月09日)

这个有标复句中, 依次出现了“是”“还是”“是”, 第一个“是”居于“还是”之前, 它们一起构成了合用型关系词, 第二个“是”居于“还是”之后, 没有跟“还是”构成合用型关系词。这是因为在“是”跟“还是”这对合用型关系词中, “是”属于“后指式”关系词, “还是”属于“前指式”关系词, 这就意味着如果某个有标复句出现了“是”, 那么我们需要考虑的是其后是否有“还是”与其呼应, 而无须考虑其前的“还是”; 反之, 如果某个有标复句出现了“还是”, 那我们要注意的是其前是否有“是”与之呼应, 而不必考虑其后的“是”。当然, “是”居于“还是”之前只是它们构成合用型关系词的必要条件, 并不是说只要“是”居于“还是”之前就一定会导致它们构成合用型关系词。但是如果“是”不居于“还是”之前, 那它们一般不会构成合用型关系词。

要人去理解其中的道理, 并不是一件很难的事, 但要计算机去理解却不是一件很容易的事。因此我们

收稿日期: 2007-08-17

基金项目: 华中师范大学“211工程”重点学科建设项目“小句中枢说在汉语与临界领域研究中的应用与验证”子项目“小句中枢理论的应用与复句信息工程”资助

作者简介: 尹蔚(1980-), 女, 湖南株洲人, 华中师范大学博士研究生, 主要研究方向: 现代汉语语法, 中文信息处理; 罗进军(1978-), 男, 华中师范大学博士研究生, 主要研究方向: 现代汉语语法, 中文信息处理。

要用形式化的语言告诉计算机如何自动识别“是”跟“还是”这类合用型关系词, 经过较为深入的考察, 我们发现“语序制约”是一条比较重要的规律: 在对诸如“是”跟“还是”这类合用型关系词进行自动识别时, 始终需要注意的是“后指式”关系词必须居于“前指式”关系词之前。有了这条规律的制约, 计算机就会准确地对诸如上例的有标复句进行自动识别。

## 二、相互制约律

哲学常识告诉我们, 客观世界中彼此联系的事物之间总是相互依赖、相互制约、相辅相成的。合用型关系词之间的关系也是如此: 一方的存在总是影响到另一方的存在, 一方的存在总是依赖于另一方的存在, 这就是我们所说的“相互制约律”。先看个例子:

(2) 吃什么亏, 我从没有‘红眼病’, 也没有攀比心理, 不论是当官, 还是做人, 要的是正道, 是问心无愧!。(《人民日报》2001年06月26日)

观察可知, 第一个“是”跟其后的“还是”既符合“语序制约律”, 也符合“相互制约律”, 它们一起构成了表选择关系的合用型关系词。而第二个“是”跟第三个“是”既不符合“语序制约律”, 也不符合“相互制约律”(其后没有“还是”与之呼应), 因此没有构成合用型关系词。

需要说明的是: 第一, 在对合用型关系词进行自动识别时, “相互制约律”的运用要以“语序制约律”为前提条件。第二, 我们就说构成合用型关系词的各个组成部分是相互制约的, 但其制约力度却不一定是完全对等的, 也就是说有可能存在这样一种情况: 一方在这种制约性关系中占据主导地位, 而另一方则处于从属地位。就“是”跟“还是”而言, “还是”占主导地位, 而“是”处于从属地位, 表现为占主导地位的“还是”有时在单用的情况下也可以标示选择关系, 但是处于从属地位的“是”单用则一般不能标示选择关系。如:

(3) 你去呢, 还是我去?

(4) 你去呢, 是我去? (×)

例(3)中的“还是”单用, 标示选择关系, 例(4)中的“是”不仅不能标示选择关系, 而且整个句子都说不通。当然, 我们并不是说“还是”在任何情况下单用都标示选择关系, 相反, “还是”在很多情况下单用都不标示选择关系。但是如果有“是”跟它配套使用, 而且遵循前面的“语序制约律”, 那么它们一般说

来会在这种相互制约下构成合用型关系词, 标示选择关系, 这样的例子不胜枚举:

(5) 究竟是能而不为, 还是为而不能? (《人民日报》1984年03月08日)

(6) 是见义勇为, 还是只看热闹? (《人民日报》1986年12月20日)

无庸讳言, 语言事实太丰富, 单凭“相互制约律”无法管控住所有的情况。但是我们同时又必须承认, 正是有了这种相互制约, 合用型关系词自动识别的准确率得到了较大的提高, 我们目前正在从事的对合用型关系词的自动识别可以证明这一点。

## 三、句法空间制约律

上文提到, 鉴于汉语的灵活性以及丰富性, 有不少含有“是”跟“还是”而且遵循了“语序制约律”和“相互制约律”的句子, 其中的“是”跟“还是”也不一定是标示选择关系的合用型关系词。如:

(7) 铜选班大部分是外来的打工仔, 也还有小部分当地工, 少不了摩擦, 他付出一份公心, 收获一份爱心, 如今, 他这个班长的交椅, 还是稳稳地坐着。(《人民日报》1993年10月26日)

上例的“是”跟“还是”尽管遵循了“语序制约律”和“相互制约律”, 但它们不是标示选择关系的合用型关系词。这是为何? 经过我们的认真考察, 发现“是”跟“还是”能否成为标示选择关系的合用型关系词, 制约的因素是多方面的, 除了语序制约和相互制约外, 还有句法空间制约。简而言之就是合用型关系词组成部分之间的句法空间越近, 他们成为合用型关系词从而标示某种关系的可能性越大, 反之则越小, 这就是我们所说的“句法空间制约律”。以“是”跟“还是”为例, 它们句法空间最近的时候可以处于同一个字段, 成为一种紧缩形式, 如:

(8) 但是, 不论是石涛的山水画还是朱牵的花鸟画, 当作中国艺术文化观察时, 我看不宜以西方抽象派艺术当作衡量它们得失的标准。(《人民日报》1996年04月11日)

不过由于我们对关系词的自动识别是为以后的工作服务的, 具体说来就是为有标复句层次关系的自动识别服务的, 而对有标复句的层次关系进行自动划分标注只能在字段与字段之间而不能在字段之内, 所以对于这种合用型关系词的紧缩形式我们将运用规则予以制约, 计算机不需要对它进行自动标注。因此,

合用型关系词的组成部分最佳的句法空间是分属两个相邻字段,也就是说这种情况下它们成为合用型关系词的几率最大。在我们所考察的19000多条含有“是”跟“还是”的语料当中,绝大多数标示选择关系的“是”跟“还是”都是分属两个相邻字段。反过来,那些句法空间间隔的较远的“是”跟“还是”成为合用型关系词的几率是比较低的。

(9) 第一层是马克思已阐明的,不管是借入资本,还是自有资本,企业总利润按照资本的单纯所有者(即马克思说的“法律上的所有者”)与资本的使用者(即马克思说的“经济上的所有者”)两重职能和权利,先分割为利息与企业主收入。(《人民日报》2000年12月14日)

例(9)中第二个字段的“是”跟第三个字段的“还是”句法空间最佳,最容易构成标示选择关系的合用型关系词,事实上也是如此。而第一个字段的“是”跟第三个字段的“还是”的句法空间相对而言较远,构成合用型关系词的几率较低,事实上也没构成合用型关系词。当然,句法空间的制约也是相对而言的,并不是说句法空间最佳的“是”跟“还是”就一定会构成标示选择关系的合用型关系词,也不是说句法空间较远的“是”跟“还是”就一定不会构成标示选择关系的合用型关系词,这是因为:第一,语言生活中的“例外”是经常有的,正是由于这些“例外”无时不在、无处不在,所以语言生活中的很多规律都体现为倾向性规律,我们很难用一个绝对的框架去管约所有的语言现象。第二,诚如上文所言,制约合用型关系词的因素是多方面的,除了以上三条较为重要的规律外,下面这条规律的制约作用也是不容小觑的。

#### 四、亲密度制约律

先看两个例子:

(10) 当年香港的丢失和今日香港的回归不仅是国力强弱的对比,还是封闭与开放的对比。(《人民日报》1997年07月05日)

(11) 即便实际上是购买了一件奢侈品,如今的人们还是希望自己看起来既休闲又酷。(《长江日报》1998年10月09日)

这两个例子中的“是”跟“还是”都遵循了前面我们所提到的三条规律,可这里的“是”跟“还是”却不是标示选择关系的合用型关系词,原因何在?仔细观察之后,我们发现:前例中的“还是”不再与其

前的“是”合用,而是与“不仅”合用,构成标示递进关系的合用型关系词;后例中的“还是”也不再与其前的“是”合用,而是与其前的“即便”合用,构成标示让步关系的合用型关系词。之所以如此,是因为“还是”跟“不仅”或者“即便”的关系比“还是”跟“是”的关系要亲密,也就是说前者的亲密度要高于后者。因此,相对“是”而言,“不仅”或者“即便”更容易跟“还是”合用,从而导致了处于这种句法环境下的“是”跟“还是”不再是标示选择关系的合用型关系词。

现在的问题是:为什么“不仅”或者“即便”跟“还是”的关系要比“是”跟“还是”的关系亲密呢?简而言之,这是一种“强式规约”。因为“不仅”也好,“即便”也罢,都属于强式关系词<sup>②</sup>,而“是”属于弱式关系词<sup>③</sup>,所以当强式关系词跟弱式关系词连用时,前者的规约力度会明显强于后者,这样便使得“还是”更倾向于跟“不仅”或者“即便”合用。这就好比“不仅”或者“即便”是块大磁铁,“是”是一块小磁铁,而“还是”则是一个小铁钉,当三者同时出现的时候,小铁钉会由于大磁铁的巨大磁场而偏向于它,小磁铁就只有望尘莫及了。但是,如果小铁钉不只一颗,那情况又将如何呢?

(12) 兵败之后,无论是声讨教练,还是发誓要抓青少年,都说明改革的力度还不到位,改革的方向不能动摇。(《人民日报》1997年11月21日)

(13) 不管是因了富而摄影,还是为了富而摄影,总之,农民摄影和农村致富是密不可分的。(《人民日报》1986年12月06日)

例(12)中,出现了“还是”跟“都”这两颗小铁钉,最终偏向“无论”这块大磁铁的是“都”,这是因为“都”跟“无论”的亲密度要高于“还是”跟“无论”,所以“无论”优先选择了跟自己最亲的“都”,构成了标示让步关系的合用型关系词。这也就是说“还是”没有受到“无论”的强式规约,因而与前面的“是”一起构成标示选择关系的合用型关系词。例(13)中的“不管”之所以选择“总之”而不选择“还是”,其原理跟前例中的“无论”选择“都”而不选择“还是”是一样的。而且,即使隐去了“都”或者“总之”这些小铁钉,“无论”或者“不管”也倾向于跟它们在时所引领的部分构成句法语义关联。如:

(14) 不管人类的未来是群体迁入太空,还是部分向太空延展,地球作为生命的摇篮,永远是人类难

以割舍的归宿,正如家园之于游子。(《长江日报》1995年12月29日)

→不管人类的未来是群体迁入太空,还是部分向太空延展,(总之,)地球作为生命的摇篮,永远是人类难以割舍的归宿,正如家园之于游子。

至此,我们可以这样界定“亲密度制约律”:假设有 $N$ 个( $N \geq 2$ )关系词能跟某个关系词合用,那么在具体的有标复句中,在其他条件等同的情况下,跟这个关系词合用的往往是亲密度最高的那个。当然,计算机本身无法判定谁跟谁更亲密,所以得事先由人提供给计算机相关信息,而这些信息的获得最好是把人的认知经验和统计方面的相关数据结合起来。

## 五、结语

复句信息化工程是一个系统工程,其中的每一个环节可能都会牵涉到很多细节问题,合用型关系词的自动识别就是如此。合用型关系词按照组成部分数量的多少可以分为“两件套”、“三件套”、“四件套”……<sup>④</sup>,不过只要我们把两件套的规律搞清楚了,其他的便可以依理类推了,这也是我们选择“是,还是”作为我们考察对象的原因之一。

如果从“强式关系词”和“弱式关系词”的角度来观察“两件套”式合用型关系词的构成,我们会发现它可以分为以下四种类型:① 强式关系词+强式关系词;② 强式关系词+弱式关系词;③ 弱式关系词+强式关系词;④ 弱式关系词+弱式关系词。其中第一种不是我们关注的重点,因为计算机对它的自动识别一般不会存在什么问题。剩下的三种是我们需要重点考察的,特别是第四种,既是重中之重,更是难中之难。有鉴于此,我们把“是,还是”作为实验对象,经过认真观察和分析之后,我们发现四条规律制约着此类合用型关系词的自动识别,它们分别是:① 语序制约律;② 相互制约律;③ 句法空间制约律;④ 亲密度制约律。

这四条规律相辅相成,共同制约着诸如“是,还是”这类合用型关系词的自动识别。必须说明的是,正如我们上文所说的,这些规律都是一些倾向性规律,我们切不可把它们绝对化,因为语言生活中“绝对化”的东西是比较少见的。尤其需要注意的是我们现在是用计算机来自动识别合用型关系词,计算机跟人脑毕竟不是一回事,它所需要的是高度形式化的东西,因

此我们得从那些外显性的东西里面找出规律,并且用形式化的语言表述出来,对那些内隐性的东西我们也要尽量使它们外显化,从而让计算机“好懂”,这样才有可能达到我们想要的“智能”。不过这事说起来容易,做起来却并不简单。这是因为:首先,我们必须尽可能全面地去观察语言现象,去体悟其中的制约机制,诚如邢福义先生所言“根据汉语特点,研究汉语句法机制必须探究‘句管控’……”<sup>[2]</sup>,正是在这样的理论背景下,我们竭尽所能地去探求制约合用型关系词自动识别的规律。其次,我们总结归纳的规律得让计算机好懂,也就是说便于程序表达。最后,我们需要一次又一次的实践论证,因为只有经过实践证明了的规律才能算是可行的规律。

我们已有的实践告诉我们:把上述规律应用于诸如此类的合用型关系词的自动识别之后,准确率有了一个比较大的提升。以“是,还是”为例,在没有把这些规律付诸实践之前,计算机对它的自动识别的准确率是很低的,可是有了这些规律制约之后,准确率就上来了,我们随机抽查了300条含有“是,还是”的语料,其中准确标注的有281条,准确率为93.67%。不过我们也知道,鉴于汉语的灵活性和复杂性,想靠有限的规律管约住所有的语言事实几乎是不可能的;但是,如果我们能够凭借从实践中总结出来的规律管控住大多数语言现象,那也是一件很有意义的事情。

### 注释:

- ① 根据邢福义先生(2001)对有标复句的关系分类,“是p,还是q”划入选择关系。
- ② 所谓“强式关系词”是指那些不受太多的句法环境制约的关系词,它们只要一出现在句子当中,一般情况下都会充当关系词,除了文中提到的“不仅”“即便”外,还有“如果”“只要”“既然”……。
- ③ 所谓“弱式关系词”是指那些受特定句法环境制约的关系词,它们只有在特定的格式当中才做关系词,比如“是”,当它跟“还是”或者“不是”等有限的词语合用时,它才充当关系词。
- ④ 这里套用服装业的术语,所谓“两件套”是指某合用型关系词由两个关系词构成,如“是,还是”“因为,所以”“如果,就”……。同理可得,“三件套”是指某合用型关系词由三个关系词构成,如“要么,要么,要么”“先,接着,然后”……。其他诸如“四件套”“五件套”都可以依理类推。

(下转 755 页)