“不是”的用法及自动处理研究

张运良[1,2]

（1. 中国科学院 声学研究所，北京 100080；2. 中国科学院研究生院，北京 100039）

摘 要： “不是”的用法判别和处理策略研究是计算机对现代汉语进行深层处理的必要内容，对句类分析系统中词汇语义模糊的消解有重要作用。本文对“不是”的各种用法及其分布情况进行了研究，并从上下文的关联、语句的复杂程度、是否属于特定问句、对语句的语义影响等方面提出了各种用法的判别和处理规则。同时，本文对这些规则进行了验证，结果表明这些规则在应用上达到了较为满意的水平。

关键词： 自然语言理解；HNC 理论；词汇语义模糊消解；概念类别；“不是”

The Study on Automatic Processing of Chinese word “BuShi”

ZHANG Yun-liang[1,2]

(1.Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080; 2.Graduate School of the CAS, Beijing 100039)

Abstract: The study on decision and processing of *“BuShi”* is very important in automatic modern Chinese processing. It will benefit disambiguation in Sentence Category Analysis System (SCAS). By context, complexity of sentence, interrogative and the affection to the sentence, the decision rules and processing rules were put forward. The validity of the rules was also tested.

Keywords: NLP; HNC theory; disambiguation; conceptual category; *“BuShi”*

1 引言

“是”字在现代汉语中使用频率较高、用法较多，是最重要的表示基本判断的词汇。HNC 句类分析系统在消解语义模糊的时候，需要对“是”字做特别的小专家系统处理。本文的研究是对句类处理中通用规则的必要补充，是“是”字小专家全面研制的重要准备工作[1-2]。“不是”作为词通常有两个义项[3]，即 1）表示否定判断；2）错误，过失。当然也有些辞书认为它只有“错误，过失”这一个义项，而在表示否定判断的时候，是一个词组[4]。此外，“不是”还存在连词的用法[5]。在 HNC 概念基元符号体系[1] [6]下，“不是”有 3 种可能的概念类别：效应(r)类、动态(v)类、语言逻辑(*l*)类。效应类充当 GBK，动态类充当 EK，语言逻辑概念类起到句间或者语义块之间的连接作用。

“不是”处理主要从“不是”上下文的关联、语句的复杂程度、是否属于特定的问句、“不是”对语句的语

基金支持：国家 973 项目“自然语言理解的交互引擎研究”（2004CB318104）和中科院声学所知识创新工程项目资助。

作者简介：张运良（1979—），男，吉林九台人，博士研究生，主要研究方向为自然语言处理、HNC 理论.

义影响等方面综合考虑。紧邻的上下文比较容易判断，包括前置的 “绝”、“无”、“如果”、“岂” 等等和后置的“吗”、“指”“很” 等等；非紧邻的包括“是”“就是”“而是” 等等。此外，上下文关联还包括“不是” 前后的语义块数量和标点符号情况。这些上下文概念是判断的激活点，激活后还需使用其它可能信息加以验证。“不是”构成语句的复杂程度可以从小句中“不是” 前后主语义块的数量来做简单的度量，当然在更加精确的语句复杂性计算中，语义块的组合顺序(语句格式)也应该考虑进来。问句的判断主要通过标点符号来判断，包括问号和叹号。“不是” 对语句的语义影响则需要去掉“不是”，再利用简单的句类知识，即句类代码和语义块数量的对应性来判断。

本文的处理原则是能够化为简单形式的尽量化为简单形式以降低计算机在进行句类分析时的复杂度和难度，具体措施包括补全 GBK，去掉冗余部分或者代之以计算机容易处理而语义不变形式等等。

本文实用 HNC 理论句类和语义块知识。本文根据文献[5]所列出的复句的种类设计了一套编码方法本文中用到复句关系表示符号如下：1)“!L218” 表示“不是……而是……” 和类似的表示明确选择的广义并列复句；2)“!L252” 表示“不是……就是……” 和类似的表示二选一的广义并列复句。本文用 f 类概念表示语气，使用双层括号表示二选一语义块。

2 “不是” 的处理策略简介

3.1 “不是” 作为 v 类概念并充当 EK（A-1）

作为 v 类概念并充当 EK 是“不是” 最常见也是最基本的用法。它的基本功能表示否定判断，从而形成基本判断句，判断的对象(DB)可能是简单的语义块，也可能是句蜕或者其它复杂构成。另外，有的“不是” 是对疑问句的回答，这种回答可能是一问一答，也可能是自问自答，通常较简短，需要根据问句恢复省略的语义块。例 1-2 为例 1-1 的处理结果。

例 1-1: 你是这个社会的主导者吗？不是。

例 1-2: 你是这个社会的主导者吗？你不是这个社会的主导者。

判别规则(A-1) “不是” 处于语句的中间或者结尾，且“不是” 前面有紧邻的 uv 类概念并且没有被判断为 *l* 类概念；或者其它判别规则都不适用。

处理规则(A-1) 如果“不是” 处于句子中间，则按照正常句类处理流程处理，如果“不是” 处于句子结尾，则必然发生了语义块的省略共享，可以根据上下文恢复省略共享的语义块。

3.2 “不是” 作为 l 类概念（A-2）

“不是” 可以不表示判断的意义，而成为语言逻辑概念，充当连接相关的语句或者语义块的功能，连接的语义块可能是主块，也可能是辅块。例 2-1 中仅仅出现了“不是” 而没有出现相互搭配的 *l* 类概念，它和“玻璃” 构成一个选择关系，处理后结果如例 2-2 所示。此外还有和 *l* 类概念“而是”、“就是” 搭配共同构成广义并列关系的情况，例略。

例 2-1: 坛上的栏杆用玻璃制成，而不是传统的汉白玉。

例 2-2: 坛上的栏杆用玻璃(f44(汉白玉))制成。

判别规则(A-2) “不是” 处于语句的开始，通常以“……[而]不是……”“是……[而]不是……” 和“不是……[而]是……”“不是……就是……” 的形式出现，并且去掉这些词对单个语句的完整性没有影响。

处理规则(A-2) 将“[而]是”“[而]不是”“就是” 去掉。如果连接的是两个语义块将两个语义块写在一起，并且为“不是” 引导的语义块添加否定标记“f42”，对“是” 引导的语义块不加标记，将“就是” 引导的语义块添加可选标记“(())”；如果连接的是两个句子，在后一个句子之前添加关系标记!L218 或!L252。

3.3 “不是” 同时作为 l 类和 v 类概念(A-3)

在“不是” 从 v 类概念向 l 类概念的转变中，还存在一些过渡状态。在这些过渡状态中“不是” 同时起着这两种概念的作用。无论哪一种概念类别，都将会影响最后的句类分析处理的结果。A-3 用法处理策略结合了 A-1

和 A-2 两种类型的处理策略，本着从简的原则，将其处理为 v 类概念的情况，但是要添加复句间的关系标记，如例 3-1 和 3-2 所示。

例 3-1: 发展小城镇不是一般的改革措施和发展步骤，而是我国经济社会发展的一个大战略。

例 3-2: 发展小城镇不是一般的改革措施和发展步骤，!L218 是我国经济社会发展的一个大战略。

判别规则(A-3) “不是”处于语句的开始或者中间，通常以“是……[而]不是……”、“不是……[而]是……”和“不是……就是……”等形式出现，并且去掉这些词语句中没有适当的词语充当 EK。

处理规则(A-3)为两个句子之间添加关系标记!L218 或!L252。如果是 L252，还需将所有的“不是”“就是”替换为“是”。

3.4“不是”的虚用（A-4）

“不是”虚用仅仅是表达了一个否定的含义，而判断蕴涵在论述之中。例 4-1 里的否定焦点可以是“适合”，也可以是“所有人”，所以就把整句作为否定的焦点，处理结果见例 4-2。尽管最简单的处理办法是把“不是”中的“是”去掉，但本文处理的目的是表示出否定的焦点，为进一步的处理和分析奠定基础。

例 4-1: 保健品并不是适合所有的人。

例 4-2: 保健品(f42(适合所有的人))。

判别规则(A-4) 除了“不是”有其它的可以充当 EK 的概念，并且“不是”可能在小句首或者小句中，并且去掉“不是”后，小句仍具有完整的语法语义结构。

处理规则(A-4) 如果去掉“不是”并且在否定焦点上标记，如果可能的否定焦点有多处，或者否定焦点无法判断，将全句中“不是”后的所有内容作为否定焦点。

3.5“是不是”虚用构成正反问句（A-5）

“不是”常和前面紧邻的“是”一起构成正反问句或表示是否两种情况的句蜕，经统计发现“是不是”虚用的超过 80%，而“是不是”充当 EK 而实用的仅占不到 20%。一个处理实例见例 5-1，处理结果见例 5-2。

例 5-1：但我也想到，关键是自己是不是真的错了。

例 5-2：但我也想到，关键是(f41f44f42 (自己真的错了))。

判别规则(A-5) “不是”前面还存在紧邻“是”字，单独构成一个小句而前面紧邻小句中不包含“是”或者虽然包含“是”但是虚用；或者小句中还有其它适当的 EK。

处理规则(A-5) 将“是不是”去掉，如果是非句蜕的还要将句子改成陈述句，如果 “是不是”单独构成小句，则为前面的小句加上正反问句标识；如果“是不是”不单独构成小句，则为本小句加上正反问句标识。

“不是”还有其它用法，按照使用频率分别编号为 B-1~B-7 以及 C-1~C-4. 限于篇幅，具体处理方法在这里不详细介绍。

3 “不是”的分布情况和判别规则的效果

我们从 HNC 语料库中选取了 3361 篇共 210 万字的《人民日报》语料进行分析和研究，其中共出现“不是”488 次，其各种类型用法的分布情况如表 1 所示，从表 1 可以看出新闻语料中有 91.9%的“不是”属于 8 种最主要的用法。

表 1 “不是”各类型在新闻体语料中的分布表

类型	A-1	A-2	A-3	A-4	A-5	B-1	B-2	B-3	B-4	B-5	B-6	B-7	C-1	C-2	C-3	C-4
数量	139	99	88	36	33	22	17	14	8	8	8	7	3	3	2	1
比例	28.5	20.3	18.0	7.4	6.8	4.5	3.5	2.9	1.6	1.6	1.6	1.4	0.6	0.6	0.4	0.2

“不是”8 种主要用法的判别规则的召回率和准确率如表 2 所示。

表 2 “不是”各种主要用法判别规则的召回率和准确率

类型	A-1	A-2	A-3	A-4	A-5	B-1	B-2	B-3

召回率	100%	97%	96.6%	97.2%	100%	86.4%	100%	92.9%
准却率	90.3%	94.1%	93.4%	92.1%	86.8%	90.5%	84.5%	81.3%

对于判别规则失效的情况进行分析，发现主要有以下两个原因：

1) 跨越语句的特征难以发现和使用，如例 6 中作为 B-7 型用法标志的“不是”和“的”在形式上被逗号分割开，可能误识别为 A-4 型。例 7 中“不是”分句和“而是”分句中间存在另外一个分句，可能导致“不是”被识别为 A-1 型，而实际上是 A-3 型。

例 6: 重大科学发现不是按常规计划，在可预见结果的情况下就可以得到的。

例 7: 《李向群》不是情节剧，不靠故事悬念吸引人，而是注重以人物在事件中表现出来的情感打动人。

2)特殊的反问句，如例 8 存在“不是”和“？”但是首先这个问句存在特征跨语句的问题而且和通常的反问句格式不一样，“不是”容易被判别为 A-3 型。例 9 处于引语中，没有问号出现，而且容易被判别为 A-1 型或 B-7 型。

例 8: 这不是同恶相济、相互渔利，又是什么呢？

例 9: 奶奶说：“不能那么讲。你懂的许多事，还不是从那些书里学的”

4 结论和下一步工作

本文介绍了“不是”的各种不同类型用法，尤其是“不是”从动词向连词过渡以及“不是”的虚实两种可能。本文也从上下文的关联、语句的复杂程度、是否属于特定的问句、“不是”对语句的语义影响等方面提出识别和处理的规则，并验证了这些识别规则的表现。

从自然语言处理的角度看，“不是”的处理策略是“是”处理策略的一个真子集，在“是”的处理中，凡是前面有“不”紧邻出现的都可以归结为“不是”的情况加以处理。进一步综合和分析“是”的其它用法，并寻找有效的处理策略，最终形成“是”处理的小专家系统，是本文的下一步目标。

参考文献:

[1] 黄曾阳. HNC(概念层次网络)理论——计算机理解语言研究的新思路[M]. 北京：清华大学出版社，1998

[2] 晋耀红. HNC(概念层次网络)语言理解技术及其应用[M]. 北京： 科学出版社，2006

[3] 罗竹风主编. 汉语大词典（第一卷）[M].. 上海：上海辞书出版社，1986.

[4] 李行健主编. 现代汉语规范词典[M].. 北京： 外语教学与研究出版社；语文出版社，2004

[5] 邢福义.汉语复句研究[M]. 北京：商务印书馆，2001

[6] 苗传江. HNC(概念层次网络)理论导论[M]. 北京：清华大学出版社，2005