

NLES 对句层“花园路径现象”的规避类型研究 ——基于 NV 互动型的探讨

杜家利, 于屏方

DU Jia-li, YU Ping-fang

鲁东大学 外国语学院/教育部汉语辞书研究中心, 山东 烟台 264025

School of Foreign Languages of Ludong University/Chinese-Based Center for Dictionary Research Directly under the National Education Ministry, Ludong University, Yantai, Shandong 264025, China

E-mail: dujiali68@yahoo.cn

DU Jia-li, YU Ping-fang. Research on category of avoidance of syntactic garden path phenomenon in natural language expert system—An explanation based on interactive Noun-Verb category. Computer Engineering and Applications, 2008, 44(25): 136-139.

Abstract: Garden Path Phenomenon can be partly analyzed in Natural Language Expert System. It is the core of analysis in this article, according to the discussion of interactive noun-verb category, check the processing procedures, the principle of inference and the formula of category when the processing breakdown happens. It is, by means of establishment of a special avoidance schema, formation of acceptable rules and stimulation of a match formula in which variable-character-series can be obtained by variable-characters limited by invariable-character-series, that this sample-based article testifies the fact NLES can be used to avoid noun-verb-related GPP to some degree.

Key words: natural language expert system; syntax; garden path phenomenon; processing breakdown

摘要:自然语言专家系统(NLES)对花园路径现象(GPP)的解读具有可释性,如何借助 NV 互动型的探讨来理解系统处理行进式错位时的工作流程、推理规则和模式算法是讨论的核心。通过构建独特的规避性推理视图,推导可解读的规则产生式,以及模拟以非变量字符串界定变量字符并得到变量字符代表的字符串的匹配模式,以实例验证了 NLES 对规避 NV 互动型 GPP 所具有的型的特征。

关键词:自然语言专家系统;句法;花园路径现象;行进式错位

DOI:10.3778/j.issn.1002-8331.2008.25.041 **文章编号:**1002-8331(2008)25-0136-04 **文献标识码:**A **中图分类号:**TP18

1 引言

专家系统(Expert System)是以计算机中专家知识及推理等技术来理解与求解问题的知识系统^[1],是某领域内具有专家理解能力的程序系统^[2]。自然语言专家系统(Natural Language Expert System)是以自然语言为知识库进行推理和求解,并能以库内知识为基础进行运算的解题程序系统。自1968年费根鲍姆等开发 DENDRAL 后,多类型系统如 PROSPECTOR, MYCIN, XCON 等相继建立。现信息输入简便易行已基本摆脱人工方式且推理趋向智能化。NLES 使用对英汉翻译中自然语言生成的正确性和健壮性有决定作用。

“花园路径现象”(Garden Path Phenomenon)在1970年由 Bever 首倡。随后, Foss & Jenkins, Kimball, Milne, 等在认知、心理、句法、语义和语境诸方面展开研究。

GPP 是语义流中途折返、另路通过的特殊现象,是认知既定模式的破旧立新;其得名于理解中的迷途知返;恰似走入园

中一条不能通达的幽径;径绝路封,须重择畅路;是信息处理中的理解休克,认知语义环路的暂时断开及对先期理解的否定^[3];其形成对认知感受省力原则、词频提取趋高原则和语义匹配瞬时原则提出挑战^[4]。

GPP 是语言学和自然语言处理的交叉,主要探讨语言能力语法(Competence Grammar)分析和句法分析(Parser)的关联性^[5]。有学者^[6-7]认为两分析无显著差异,但 Pritchett 对 GPP “行进式错位”(Processing Breakdown)^[8]研究后认为两者绝然不同但有关联,且涉及到局部歧义^[9]。

NLES 处理 GPP 时,秉承顺次原则,由关键词、短语再到整句匹配。词匹配效率高但易屏蔽非关键词信息;句匹配精度高但句长与效率负相关^[9]。相关研究^[10-11]认为产生 GPP 原因很多,关键词词性变化属其一。名(Noun)、动(Verb)转换所致行进式错位属 GPP 的 NV 互动型。本文倡导“立足词兼顾句”的匹配原则以分析 NLES 规避 NV 互动型 GPP 时的类型特征。

基金项目:教育部项目(No.06JJD740007);鲁东大学项目(No.W20072602)。

作者简介:杜家利(1971-),男,硕士,研究方向:篇章语义学和计算语言学;于屏方(1971-),女,博士,研究方向:应用语言学和计算语言学。

收稿日期:2007-12-05 **修回日期:**2008-02-25

2 NV 互动型 GPP 的界定

GPP 是特殊认知现象,是意识流的语义短路;层级存在性、理解折返性、认知顺序性和语义纠错性是其基本特征。其产生不局限于句层,本文在此层面对其进行讨论。

GPP 产生是多学科交叉的结果;其中,Collins 和 Quillian 所倡导的认知节俭特性,即知识库中节点最近内容首先被提取,并尽快构建语义环路以实现认知瞬时理解的特性,是 GPP 产生的认知心理学基础。

NV 互动型 GPP 是 NLES 进行英汉翻译时,对具有名词和动词双重释义的词产生的行进行式理解错位;其匹配模式兼跨词句;对其理解的折返带来认知顿悟和归位;认知提取顺序性决定靠近知识库核心的原型名词或动词(本文只讨论原型名词)首先被提取,并瞬时构成语义环路,直至理解停顿并启动语义纠错功能。对 NV 型的翻译需要打破认知节俭特性,选择认知感受费力、词频提取较低和语义匹配非常规的翻译选项。请见下例

例 1: The old man the oars.

对例 1 的常规译法呈线性模式,即按照 Kimball^[12]封闭原则(The Principle of Closure)首先确定[The old man]NP+[the oars]NP 模式,但该匹配不符合成句的语法条件,原语义环路断开,迫使语义流返回到句首重新匹配,GPP 产生。

关键词 man 具有名动两种释义,其中名词性释义是认知原型,词频较高而且具有语义匹配的常规性特点。GPP 出现后,其名词性释义向动词性释义的转换,带来句子模式由 NP+NP 向[The old]NP+[man]V+[the oars]NP]VP 的调整,系统翻译完毕。

例 1 译文:老人在操桨。

NLES 在由非自动处理向自动处理转化的过程中,需规避 NV 互动型 GPP,方能提高系统的准确性和智能性。

3 NLES 对 NV 互动型 GPP 的规避性推理

推理就是按照特定规则、理论和策略,从已知判断或结论推导出未知判断或结论、对自然语言求证解答的过程,通常包括多种策略和推理方法。规避性推理特指对某些类型的防范性判断或结论,旨在加强系统的前瞻性,减少系统的回溯性,提高系统的能率性。在 NLES 中,规避性推理基于知识库中的既定信息展开,并经过传递、匹配、消歧、限定和求解等自然处理策略,采用单元、二元和多元方法进行推理,最终实现对某些问题合乎逻辑的判断和解答。

自然语言文本生成主要有模板、Schema、修辞结构理论(RST)和基于特征 4 类技术模式,是研究计算机对文本信息的程序化表示,以及如何推理生成高质量的自然语言文本,并且在规则库映射下根据外部变化而做出相应调整的处理过程。借鉴 NLG 系统的技术视图模式^[13]和语义排歧流程图^[14]构建一个“NLES 对 GPP 规避性推理视图”,以便于理解其推理过程见图 1。

3.1 非 GPP 句法生成式研究

信息首先从“信息传递”进行输入,并在“内程序知识库”(即计算机程序模块)和“外程序知识库”(即程序外认知知识)监控下进行“词/短语瞬时匹配”,并将匹配后信息传到“整句匹配”;如符合“语法”,直达“句法生成式”并接受“语义监控”;语义和谐,便成功输出。否则,语义环路断开,推理机制启动,理解

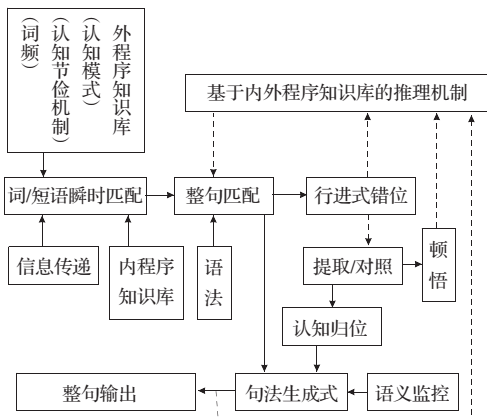


图 1 NLES 对 GPP 规避性推理视图

重新返回到“整句匹配”,并再次接受语法和语义监控,直至正确输出或系统提示不可输出。

这一过程强调语法和语义的先后性,请见下例:

例 2: We knew that when that snake bit you, you died.

(我们知道当你被那蛇咬的时候,你就完了。)

例 2 符合语法和语义,在 NLES 中翻译较为便捷,不会出现回溯。

对符合语法、能够到达“句法生成式”的句子,会因语义不符,而重新处理。

例 3: We knew that when you bit that snake, it died. (我们知道当那蛇被你咬的时候,它就完了。)

例 3 符合语法,但因语义不符难以形成环路,只能重返进行整句匹配,直到系统提示本句是不能输出的错句。本文不讨论上述类型,而主要关注产生行进行式错位的 GPP 句式。

3.2 GPP 句法生成式研究

形成“行进行式错位”的 GPP 类型有很多,如下例:

例 4: The novice accepted the deal before he had a chance to check his finances, which put him in a state of conflict when he realized he had a straight flush.^[15]

理解时,“novice-deal-finance-conflict”形成语义环路(即经济类的“新手-交易-财政-冲突”),且在“语法”监控下进入“句法生成式”。但至“a straight flush”¹时,原环路断裂,迫使语义流返回接受推理、匹配,并进入“行进行式错位”阶段,对存在于推理机中的多种模式进行“提取/对照”。不能到达“认知归位”的匹配模式由“顿悟”重返推理,直至最后匹配为娱乐类的“新手-发牌-底牌-矛盾”,实现“认知归位”并通过“语义监控”的“句法生成式”,完成“整句输出”。

例 4 译文:这个(牌局)新手在还没有检查底牌的情况下,就同意发牌了,当他意识到手中的牌是“同花顺”时,陷入了矛盾之中。

本文不讨论此类系统较难处理的 GPP,而侧重相对简单的 NV 互动型 GPP。请见下例:

例 5: The new record the music.

对其理解,经历了一次瞬时语义环路的断开。从“信息传递”而来的语义流,在知识库监控下,完成“词/短语瞬时匹配”后进入“整句匹配”阶段。因瞬时语义流不符合语法,而进入“行进行式错位”,并启动推理机制,重返“整句匹配”进行再组合。

¹a straight flush: 连续的 5 个数字,而且是同一种花纹,这是扑克中最好的牌,俗称“同花顺”。

以 record 名词释义为核心的 The new record 作为认知原型的 NP 先进入“整句匹配”, the music 与之形成 NP+NP 模式, 接受“语法”监控; 不符, 遂进入“行进式错位”, 并通过推理机制返回再匹配。最后, 由名词调整为动词释义的 record 匹配方式 NP+[V+NP]VP 符合语法, 遂在“语义监控”下经“句法生成式”完成“整句输出”。

例 5 译文: 新人在录制音乐。

此类 NV 互动型 GPP 推理, 可用规则解读。规则可视为事实的伸延, 只是附加了条件。符合条件规则才为真。其实质是储存的查询。NLES 推理规则是人工定义的, 其集合形成规则库^[18]。产生式如下(规则以逗号分隔):

[T1]record (All)_record (Noun) (X, Y): -record (Noun) (Y, X), Polysemy (X)

record (All)_record (Noun) (X, Y) 表示 X, Y 两者是上下语义关系(hyponymy)²或总分关系(meronymy), record (Noun) (Y, X) 表示 Y 是 X 的下义词或部分词, Polysemy (X) 表示 X 是多义词。[T1]表示: 如果 Y 是 X 的下义词或部分词, 并且 X 是多义词, 那么 X 和 Y 就是上下语义关系或总分关系。

[T2]record (All)_record (Verb) (X, Y): -record (Verb) (Y, X), Polysemy (X)

同理, [T2]表示: 如果 Y 是 X 的下义词或部分词, 并且 X 是多义词, 那么 X 和 Y 就是上下语义关系或总分关系。

名词释义的 record (Noun) 和动词释义的 record (Verb) 同属 record (All) 的同级下义(co-hyponymy)或部件关系(co-meronymy); 当系统处理发生“行进式错位”时, 同级关系的释义将作为首选替换, 完成系统解答。

4 NLES 模式匹配法则对 NV 互动型 GPP 的能释性

模式匹配分为内程序和外程序匹配两种。内程序匹配是指将生成模式和计算机内部程序进行比较和耦合, 检查两者的相似度, 并根据相似度比率由高至低给出匹配的顺次。外程序匹配是指生成模式和计算机外部的认知模式进行对照, 并检索出符合认知领域的多种模式, 按照匹配的完美性和顺畅性依次排列。

模式匹配是推理机制存在的关键, 只有经过模式匹配的知识库中知识才能推理。就内程序匹配的关键元来说, 现行的模式通常分为关键字为主的单元匹配和词句为主的多元匹配。

单元匹配以预存的关键词为基准构建单一或多模式, 并与一个或多个解释相对应。当系统接受处理命令时, 潜存模式被顺次提取与特定句子中关键词进行对照, 相似度在合理范围内即可得到句子的解释。此方法系统处理简单、顺畅度高、准确性好而且效率力强, 不足在于系统不能自主决定关键词, 对程序外的重要词成分无条件过滤, 容易屏蔽关键词外的重要信息。

多元匹配秉承句法原则, 强调句子和系统内存句法模式的一致性。系统在处理输入的句子时, 严格按照句法提取和对照。由于句子变化的多样性和能动性, 导致系统处理必须拥有海量信息, 而且句法模式必须达到巨量, 输入的句子必须符合句法模式才能得到一个较为满意的系统解答, 否则将导致系统理解效率低下。此项优势在于对句中涉及的内容具有涵括性, 不至于丢失核心内容。

本文讨论的模式匹配介于两者之间, 即侧重单元匹配能效性的同时兼顾多元匹配的涵括性。在现行模式基础上, 提出对 NV 互动型 GPP 具有能释性的模式匹配。

先设定(1)传递到系统的句子 IS(Input Sentence)中含有名动释义的关键词 KW(Key Word); (2)由 KW 自由组成的名词释义字符串为 NBCS (Noun-Based Character Series); (3)由 KW 行进式错位后组成的动词释义字符串为 VBCS (Verb-Based Character Series); (4)内程序模式为 IP (Inner Pattern)。

如能通过算法验证 IP 与 KW 的正确匹配, 且能兼顾 KW 外的非关键词, 则表示此模式对 NV 互动型 GPP 具有能释性。算法如下:

(1)确定 IP 中所有非变量字符串及其位置、变量字符及其位置, 转(2);

(2)利用 IP 中所有非变量字符串顺序匹配 NBCS, 匹配位置开始为零, 根据每次匹配的非变量字符串的长度依次后移, 如果匹配失败, 错误返回; 否则转(4);

(3)接(4)返回的错误匹配后, 利用 IP 中所有非变量字符串重新顺序匹配 VBCS, 匹配位置重新清零, 根据每次匹配的非变量字符串的长度重新后移, 如果匹配失败, 错误返回; 否则转(5);

(4)后移长度足以验证 NBCS 的释义错误时, 返回到(3), 否则转(5);

(5)确定 IP 每个变量表示的 KW 中字符串, 转(6);

(6)如果 IP 中的一个或多个变量表示的字符串为空, 错误返回; 否则转(7);

(7)如果 KW 同时匹配到多个模式, 取变量个数最多的模式作为最佳模式;

(8)正确返回。

此算法兼顾了能效性和涵括性, 采用非变量字符串界定变量字符, 得到变量字符代表的字符串, 融合了非变量的关键词字符串和变量的非关键词字符串, 是匹配模式的一个进步, 对认知模式由名词性释义转变为动词性释义的 NV 互动型 GPP 具有能释性。

5 NLES 对 NV 互动型 GPP 的实例解答

基于对 NV 互动型 GPP 规避性推理和模式匹配的分析, 模拟一个 NV 互动型 GPP 问答系统, 允许用户通过自然语言, 询问该类型花园路径句的翻译。

除信息输入和输出须具备的人机界面外, 系统结构(如图 1)包括 4 个子系统: (1)内程序知识库(包括结合模式匹配机制能完成知识转换任务的模板库); (2)语法(主要基于句法); (3)基于内外程序知识库的推理(包括结合 prolog 可完成推理机制的规则库); (4)语义监控(用来监控语义表达式的可受性)。

以“The new record the music.”为例解释系统工作流程:

(1)通过内程序知识库子系统, 将例句知识转换为计算机内部形式, 并存储。如本句中除定冠词 the 之外, 共有三个可存储的知识节点, 即 new, record 和 music。以 LDOCE (Longman Dictionary of Contemporary English) (2003 版)释义为来源库进行库容信息输入。

new 只有形容词释义, 在词典中除词组举例外共有 10 条

²按照词汇语义学概念, 上下语义关系或总分关系是两种不同的关系, 如“车/红旗轿车”属于上下义关系而“车/车轮”属于总分关系, 此处不区分两种术语关系。

义项。record 具有名动释义,含名词释义 4 条,动词释义 3 条。music 名词释义有 3 条。由词典输入到库容信息知识将成为系统解答的依据。预处理完毕后转(2)。

(2)通过人机界面输入“如何翻译句子‘The new record the music’?”转(3)。

(3)模板库对问题进行转换,得到形式化表示。Kimball 的“自上而下原则”(Top-Down Principle)决定定冠词 DET (determiner)首先与名词 N 结合成 NP(noun phrase)形式来对后续成分进行预测和扩充,介于 DET 与 N 之间的成分作为补充成分 AP 存在。[[the]DET+[new]AP+[record]N]NP+[[the]DET+[music]N]NP 模式首先匹配成功,new 的 10 条形容词释义与 record 的 4 条名词释义组成翻译集合如下:

表 1 new 形容词释义与 record 名词释义的翻译集合表

N	R			
	INFORMA- TION	HIGHEST/ BEST EVER	MUSIC	PAST ACTIVITIES
RECENTLY MADE	1	2	3	4
RECENTLY BOUGHT	5	6	7	8
NOT THERE BEFORE	9	10	11	12
NOT USED BEFORE	13	14	15	16
UNFAMILIAR	17	18	19	20
RECENTLY ARRIVED	21	22	23	24
RECENTLY CHANGED	25	26	27	28
RECENTLY DISCOVERED	29	30	31	32
MODERN	33	34	35	36
VEGETABLES	37	38	39	40

如表 1,两者可组成 40 个翻译单元,通过系统优化和“外程序知识库”匹配,认知不符的(非方框单元)错误返回;否则转(5)。

(4)选择 record 名词释义为匹配模式,转(5);否则转(8)。

(5)转入的 23 个单元进入“整句匹配”并和后半部的 the music 组成翻译体,限定了翻译单元的域(必须和 music 相关);与 music 义项不相关的单元错误返回;否则转(6)。

(6)过滤后的 7 个单元与 the music 组成新的翻译集合(见表 2),无语境支持的(非方框单元)错误返回;否则转(7)。

表 2 record 名词释义域内单元限定的终端翻译集合表

record 名词释义域内单元	终端翻译集合
RECENTLY MADE MUSIC	近期制作的音乐
RECENTLY BOUGHT MUSIC	近期买到的音乐
MUSIC NOT THERE BEFORE	刚产生的音乐
MUSIC NOT USED BEFORE	未曾使用的音乐
UNFAMILIAR MUSIC	不熟悉的音乐
RECENTLY ARRIVED MUSIC	刚到来的音乐
MODERN MUSIC	现代音乐

(7)从终端翻译集合中优选的 4 种翻译,均属于 NP+NP 模式,不符合“语法”,产生“行进式错位”,返回到(4)。

(8)基于知识库推理,已存储的 3 节点中只有 record 具有释义的可选择性,系统不能重新匹配为动词释义,错误返回;否则转(9)。

(9)根据 Kimball “封闭原则”(Principle of Closure),以 record 三条动词释义为核心的匹配模式首先吸纳 new 为状语修饰成分(PP),组成[new]PP+[record]V 模式,转(11);否则转(10)。

(10)record 匹配模式发生调整,与[[the]DET+[new]AP]NP 组成 NP+V 模式(即 the new 集合名词与 record 动词释义组合的模式),获得翻译集合(见表 3)后转(12)。

表 3 the new 集合名词与 record 动词释义翻译集合表

NP	V	动词释义	翻译集合
The new record		to write information down(Vt)	新人在记录+?
		to store music on tape or discs(Vt)	新人在录制+?
		to store music on tape or discs(Vi)	新人在录制。
		to measure information(Vt)	新人在测量+?

(11)new 形容词释义的唯一性决定 PP+V 模式匹配错误,返回到(10)。

(12)进入“整句匹配”阶段的 4 个翻译单元,不符合语法的单元(不及物动词 Vi 与 the music 的匹配),错误返回;否则转(13)。

表 4 record 动词释义域内单元限定的终端翻译集合表

record 动词释义域内单元	终端翻译集合
to write information down	新人在记录音乐
to store music on tape or discs	新人在录制音乐
to measure information	新人在测量音乐

(13)滤后的 3 个单元到达“句法生成式”,并接受“语义监控”(见表 4),不符合语义(非方框单元),错误返回;否则转(14)。

(14)以自然语言形式,将“he new record the music”的翻译结果“新人在录制音乐”提交给用户。

(15)正确退出。

6 结论

NLES 的应用日趋便捷和智能化,其对某些句式的翻译已具有型的特征。GPP 是行进式理解错位,其对既定认知模式的破旧立新产生系统回溯。NV 互动型 GPP 是由名动双重释义词产生的理解错位;原型释义先被提取,并构成瞬时语义环路;对其理解的折返带来顿悟和归位,并打破认知节俭特性。NLES 对此类 GPP 的规避提高了系统的精确性、健壮性和智能性。

本文构建了 NLES 对 GPP 规避性推理视图,划定了系统流程,即“信息传递”,“词/短语瞬时匹配”,“整句匹配”,“行进式错位”,“基于内外程序知识库的推理机制”,“整句(重新)匹配”,“句法生成式”和“整句输出”。提出了可进行解读的规则产生式[T1]record (All)_record (Noun) (X, Y): -record (Noun) (Y, X), Polysemy (X) 和 [T2] record (All)_record (Verb) (X, Y): -record (Verb) (Y, X), Polysemy (X)。关键词的名动释义规则说明同级下义或部件关系具有替换优先权。举例验证了模式匹配算法的能释性,即采用非变量字符串界定变量字符并得到变量字符代表的字符串的方式;兼顾了单元匹配的能效性和多元匹配的涵括性。

不足在于(1)对多学科交叉产生的 GPP,尚难有一致认定,规避性推理尚处于定性分析阶段;(2)认知复杂性决定了匹配顺序的非验证性;(3)规则产生式仅对具有名动双释义的关键词有效,多释义关键词或非 NV 互动型 GPP,无法确定替换优先权;(4)算法虽考虑了关键词和非关键词(变量字符),但变量字符必须由非变量字符串来确定的局限性决定了对变量字符比邻情况解释的无效性。

(下转 232 页)