从迭句中辨识出三类花园幽径句 *

池哲洁[1,2]，池毓焕[2]，张全[2]

(1.中国科学院研究生院 北京 100039; 2.中国科学院声学研究所 北京 100190)

E-mail: chizhejie@sina.com

摘 要：在大句的范围内小句的组织结构会呈现某些特定的模式，即大句范式。而范式的运用存在着语种间的有无或常用罕用之别，需要在翻译时予以变换。迭句与花园幽径句都是汉语的常用大句范式，而且二者容易混淆。本文着重探讨在汉英机器翻译时如何把花园幽径句从无头迭句中识别出来，给出具体的辨识算法，并提出相应的汉英转换规则，在实验结果部分分析了该算法的不足之处和改进之道。

关键词：大句范式，迭句，花园幽径句，汉英机器翻译

Identifying Three Types of Garden-Path Sentence from Main-chunk-overlapped Sentences

Chi Zhejie[1,2] Chi Yuhuan[2] Zhang Quan[2]

(1. Graduate School of the Chinese Academy of Sciences, Beijing 100039, China ;

2. Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China)

E-mail: chizhejie@sina.com

Abstract: In the viewpoint of Major Sentence, there should be some certain modes on organizing its minor sentences that are named as Paradigm of Major Sentence. There are some remarkable differences of using those paradigms of major sentence between different languages. So these paradigms should be transformed during translating. Main-chunk-overlapped sentences and garden path sentence are frequently used in Chinese, thus they are easy to be confused. In this paper, we present a method used for identifying garden path sentence from main-chunk-overlapped sentences in Chinese-English machine translation, in addition to the identification algorithm, we design a number of Chinese-English transformation rules for machine translation. In our experiments, we empirically analyze the problems with the algorithm and present some feasible solutions.

Keywords: Paradigm of Major Sentence; Main-chunk-overlapped Sentences; Garden path sentence; Chinese-English Machine Translation

1 引言

汉语中的花园幽径句，如单句：

小王研究鲁迅的文章发表了。

该句中，“小王研究鲁迅的文章”就已经是一个完整的“主谓宾”结构，这在机器识别时就会被当成一个整体来处理，而事实上，这一部分也可作为偏正结构的名词词组。在这一小句中，“发表了”的出现也就说明“小王研究鲁迅的文章”这部分应视为后者。对于这类句子，在整个识别过程中，起初会有捉摸不定之感，直到看完整个句子才恍然大悟，最终确定句子的结构。该过程有一个形象的比喻：当我们走进一个风景如画的花园，要寻找这个花园的出口，大多数人都

* 本文承 中科院声学所知识创新工程项目“句群理解处理理论及其应用”（O654091431） 的资助。

认为出口一定应该在花园的主要路径的末端，于是我们沿着花园中的主要路径欣赏花园中的美景，突然发现这条主要路径并不通向花园的出口，而能够通向花园出口的正确的路径，却是在主要路径旁边的另一条被几乎游人遗忘的毫不起眼的荒僻的幽径。以此故，把 garden-path sentence 译作“花园幽径句”[1]。

我们聚焦于潜在的歧义结构部分以独立小句形式出现的花园幽径句，如：

香港、澳门胜利回归祖国，实现了全民族的夙愿。

这意味着要以大句为语句处理单位，而不能仅以逗号截止。因此特引入以下若干概念：以句号或与其等价的问号、感叹号等为结束标志的文本片段简称语段；语段内如有逗号等分割标志，则称该语段由若干语串构成。大约 65%的语串成句[2]。如果这些语句仅是语段的构件，则称之为小句；相应地，由若干小句构成的语段称作大句。

在黄曾阳[3-4]先生引进的表征大句句式结构的“大句翻译范式[5]”中，也存在花园幽径句的句式。其形式结构混同于迭句，特别是无头迭句，而语义结构却迥异。在翻译时，若按普通迭句处理将导致严重错误，因为花园幽径句总是包含潜在的歧义结构，机器所能优先识别的歧义结构往往是错误的，故翻译前应先将其正确识别出来，翻译时进行相应的大句范式转换，从而提高翻译的正确率。

对现有的基于规则的机器翻译系统进行测试可以发现，大多数系统对此类句式只能部分识别，即与‘是’字句一同处理时可得到部分正确结果，单项处理正确率不及 50%。其中一个系统知道一、两个花园幽径句向特殊句式转换，对再多的花园幽径句就无能为力了；还有一个系统则知道在‘是’前加 it，但不知道为此前的系列小句“安头”并单独成大句（即改逗号为句号，所加的 it 首字母大写）。相比之下，对无头迭句的认识比较准确，但处理不到位，绝大多数不知“安头”，或不知在最后一个小句前加 and 以接应[5]。

由上述测试结果可知，现有的翻译系统并没有将此类句式作为统一的语言现象加以辨识并予以统一的处理，以致遇到该类句式时错误频出，不但造成部分迭句的翻译结果可读性不强，而且误将花园幽径句当迭句翻译时歪曲原意，乃至不知所云。以下给出此类特殊句式的一个实例，同时也摘出各常用翻译系统的错误处理结果。

例：坚持科学态度，大胆进行探索，使我们的思想和行动更加符合客观实际，更加符合社会主义初级阶段的国情和时代发展的要求。

系统 1 翻译（未能识别出该句式，组织结构错误）：Adhere to the scientific attitude, explore boldly, make our thought and action accord with the objective reality further, accord with the demands for development of national conditions and era of the primary stage of socialism further.

系统 2 翻译（与系统 1 类似）：Adhere to the scientism, courageously explore, make our thought and action accord with objective reality more, accord with t he n ational conditions o f p rimary stage o f socialism and t he r equirement o f era development more.

系统 3 翻译（按“使”字句翻译，可读性稍强）：Adhere to the scientific approach and make bold explorations in order that we think and act more in line with the objective reality, the primary stage of socialism and the national development of the times.

针对现有翻译系统对这些常见句式处理效果极不理想的状况，本文着重探讨如何从迭句中辨识出花园幽径句，提出识别算法，使用已有词语知识库实现该算法，同时对一定数量的真实文本进行测试，得到相应的结果，以图对现有系统有所助益。

2 句式分析

我们先从迭句的具体句式说起，对相应的例子进行分析。该分析基于现有词语知识库（其中概念类别项用于结构块的标记，如对应概念类别以“v”开头刚识别结构块为 EK 块；句类代码项用于识别广义作用句和广义效应句），仅根据各句式识别的需要给出例句的部分标注。

2.1 迭句

2.1.1 普通迭句

例 1. 我们/GBK1 彻底结束了/v 中国近代以来屈辱外交的历史，有力地维护了/v 国家的主权、安全和民族尊严。

例 2. 鲁滨逊夫人/GBK1 赞同/v 江泽民关于反对恐怖主义问题的论述，认为/v 国际社会应对阿富汗目前存在的人道主义危机予以足够的关注。

2.1.2 无头迭句

例 3. 坚持实施/v 可持续发展战略，正确处理/v 经济发展同人口、资源、环境的关系，改善/v 生态环境和美化生活环境，改善/v 公共设施和社会福利设施。

例 4. 要结合/v 形势的发展，紧紧围绕/v 党的中心任务，不断加强/v 党的建设。

以上四个例子中，各大句全部由!310 格式[6]或包含!310 格式的系列小句所组成，这就是汉语中常见的迭句。头两个例子，大句中除第一小句外，各小句均是以谓语结构 EK 打头，而第一小句的格式为 GBK1+!310EJ，且该 GBK1 是其后各小句的公共 GBK1，此大句符合 GBK1+$\{!310EJ\}_n$ 的格式，这样的句子我们称为普通迭句，对它们进行翻译所采用的转换规则为：零转换，即第一小句的主语作为整个大句的主语，其余部分按正常并列句翻译即可。后两个例子中，各小句的打头部分均是谓语结构 EK，符合$\{!310EJ\}_n$的格式，且最后一小句的结构形式与之前各小句相同，总体地位平等，打头动词并非“是”或“有”，此类句式我们称为无头迭句，翻译时的转换规则为：在第一小句安头，作为整个大句的主语，其余各部分无需进行特殊转换，按正常并列句翻译即可（并列句的最后一句前补 and 以接应）；另外，对于主语不言自明的可按英语被动式翻译。

除普通迭句和无头迭句外，由!310 格式的系列小句所组成的大句还有其他类型，可采用如下形式化描述：[GBK1]{!310EJ}|[f84EJ]（“|”表示该项可重复，中括号表示其中的项目为可选），这当中还包含了花园幽径句。

大句中的花园幽径句是由一系列!310 格式系列小句打头，这些打头小句作为整体，是最后一个小句的 GBK1，形式化描述为$\{!310EIJ\}_{n-k}$+!310EgJ 或$\{!310EIJ\}_{n-1}$+ f84EgJ。此次识别的花园幽径句又可细分三类：如果一系列!310 格式的小句打头阵，然后出现一个以重复指代 f84 之捆绑词语（如“这”和“那”）打头的小句，我们把这种大句称作花园幽径句 A，形式化描述为：$\{!310EIJ\}_{n-1}$+(f84)EgJ；相应地，以“是/有”为 EgK 的大句称作花园幽径句 B，形式化描述为：$\{!310EIJ\}_{n-1}$+!310jDJ；另外，以系列广义作用句打头加广义效应句的句式或相反形式的句式称为花园幽径句 C，形式化描述为：$\{!310E_{[X]}J\}_{n-1}$+$!310E_{[Y]}J$ 或$\{!310E_{[Y]}J\}_{n-1}$+$!310E_{[X]}J$。

以下部分将对花园幽径句进行分析。

2.2 花园幽径句A

例 5. 一方面，反对/v 一切丧尽天良的坏蛋，反对/v 那些投降派和反共顽固派，这/f84 是/v 又一条政策。

例 6. 敏锐地把握/v 我国社会生产力的发展趋势和要求，坚持/v 以经济建设为中心，不断促进/v 先进生产力的发展，这/f84 是/v 我们党始终站在时代前列、保持先进性的根本体现和根本要求。。

以上两例，大句中除最后一小句外的各小句均以谓语结构 EK 打头，最后一小句的打头部分不是 EK 结构，而是“这+是”的形式，整个大句符合“$\{!310EIJ\}_{n-1}$+这/那+是/有”的格式，故判其为花园幽径句 A。此类句式大部分与无头迭句相似，仅是最后一小句中打头词“这/那”的出现改变了句式的归属。本句式的转换规则为：将系列原型句蜕的 EI 提升为 Eg，该安头的先安头；最后一句翻译成独立的语句。

2.3 花园幽径句B

例 7. 坚持/v 什么样的文化方向，推动/v 建设什么样的文化，是/v 一个政党在思想上精神上的一面旗帜。

例 8. 在新的世纪，继续推进/v 现代化建设，完成/v 祖国统一大业，维护/v 世界和平与促进共同发展，是/v 我们党肩负的重大历史任务。

以上两例，各小句均以谓语结构 EK 打头，整个大句符合$\{!310EIJ\}_n$的格式，但最后一小句较之前几句稍为特殊，以“是”打头引领小句，整体呈现“$\{!310EJ\}_{n-1}$+jDJ/jD1J”结构，故判其为花园幽径句 B。此类句式中，最后一小句的 GBK1 是由之前各小句共同构成的。转换规则：向英语的特定句式[7-8]转换，即引导词 it 作形式主语先翻主句，按常规的原型句蜕变换规则[9]处理打头的若干并列小句，并把系列不定式短语或现在分词短语后置。当然这不是唯一可行的变换方式。

2.4 花园幽径句C

例 9. 要/v[X]相互尊重与平等互利，不要/v[X]霸权主义和强权政治，要/v[X]对话与合作，不要/v[X]对抗与冲突，已成为/v[Y]越来越多国家的共识。

例 10. 加强/v[Y]有说服力的思想政治工作，发展/v[Y]教育科技事业，繁荣/v[Y]社会主义文化，使/v[X]人人都有受教育的机会和享受文化成果的充分权利，使/v[X]人们的精神世界更加充实、文化生活更加丰富多彩。

以上两例，各小句仍以谓语结构 EK 打头，最后一小句的引领词并非“是”或“有”，整个句式呈现$\{!310EIJ\}_n$的格式，但其与无头迭句中各小句地位平等、平起平坐的特点却不相同。例 9 中前几个小句的打头结构均为句类代码为广义作用句的词，最后一小句打头结构的句类代码为广义效应句，整个大句的格式[10]为：$\{!310E_{[X]}J\}_{n-1}$+$!310E_{[Y]}J$，其中，最后一小句的出现对先前的各判断句一一加以总结，使之前的各系列小句一同降格；例 10 反之，它是各系列小句描述广义效应的各侧面，接着来一个基本判断句，格式为：$\{!310E_{[Y]}J\}_{n-1}$+$!310E_{[X]}J$。故判这两句均为花园幽径句 C，转换规则同花园幽径 B。

对第一小句查看是否含有 GBK1 可区分普通迭句与无头迭句类（包括无头迭句和花园幽径句）。关于如何从无头迭句中区分出花园幽径句，则可使用黄曾阳先生所拟的一个判别口诀：“‘是’字强出头，‘有’字带呼应；两者不存在，‘这’‘那’拿来顶；‘这’‘是’若联用，Eg 可铁定”。如此，可以说辨识特征简单明了。

3 识别算法

对于迭句（普通迭句与无头迭句）和三类花园幽径句的整体识别，首先要以大句作为输入单位，单独小句则不作为处理对象。识别预处理阶段的工作包括分词或标出成词单位在句子中的位置，切分好的字词是本次处理的基本单位；同时，需要获得各字词单位所对应的概念类别符号

与句类代码符号。完成预处理，获得上述三类信息后可进入句式识别阶段，最终输出句式的判定结果，同时在文字上给出翻译时所要采用的转换规则，但整个算法中并未实现句式的转换。

句式识别算法描述：

(1) 判断输入句子是否为大句，即至少包含两个小句的句子；若是，则进行分词并获取概念类别和句类代码信息，转步骤(2)继续，否则，结束判断，输出“非考虑句式”。

(2) 判断整个大句除头尾两句外是否都符合!310EJ 格式，即判断大句的格式是否为{?}+{!310EJ}$_{2\cdots n-1}$+{?}；若是，转(3)，否则，结束判断，输出“非考虑句式”。其中，{?}表示该句的格式有待识别；对于仅含两小句的大句，可直接进入(3)的判断。

(3) 判断第一小句是否符合!310EJ 格式；若是，转(6)，否则转(4)。

(4) 判断第一小是否符合 GBK1+!310EJ 格式；若是，转(5)，否则，结束判断，输出“非考虑句式”。

(5) 判断最后一小句是否为!310EJ 格式；若是，则整个大句格式为 GBK1+{!310EJ}$_n$，输出“普通迭句”的结果，结束判断，否则，输出“非考虑句式”，结束判断。

(6) 此时，整个大句符合{!310EJ} $_{n-1}$+{?}格式，则判断最后一小句；若最后一小句格式为!310EJ，则转(7)，否则，判断其是否为“这/那+是/有 J”格式，若是，即整个大句的格式为{!310EJ}$_{n-1}$+(f84)EJ，则输出“花园幽径句 A”，结束判断，若不是，则输出“非考虑句式”，结束判断。

(7) 判断最后一小句的打头词是否为“是”或“有”；若是，该大句符合{!310EJ} $_{n-1}$+是/有 J 格式，则输出“花园幽径句 B”，结束判断，否则，转入(8)的判断。

(8) 判断整个大句是否为{!310E$_{[X]}$J}$_{n-1}$+!310E$_{[Y]}$J 或{!310E$_{[Y]}$J}$_{n-1}$+!310E$_{[X]}$J 格式；若是，则输出“花园幽径句 C”，结束判断，否则，输出“无头迭句”，结束判断。

注：!310EJ 格式的判别方法为：小句的居首成分为谓语结构 EK，且 EK 后非空。初步处理时可仅考虑单个动词即为 EK。

上述算法中，分词及句式判别所包含的各方法的实现必须基于含有概念类别及句类代码的词语知识库，不同人所使用的知识库在构造选词方面会存在差异，实现各方法的程序也会不同，故在此不给出具体实现代码。

4 实验结果

4.1 识别算法测试结果

本次判别算法中使用的词语知识库容量大约为 4 万词（包含单字词），测试语料来源于网络上的文章，选取《新民主主义论》、《在延安文艺座谈会上的讲话》、《在庆祝中国共产党成立 80 周年大会上的讲话》、《国家科学技术奖励条例》、《中华人民共和国和俄罗斯联邦关于世界多极化和建立国际新秩序的联合声明》、《团结一切抗日力量，反对反共顽固派》和《江泽民会见联合国人权高专罗滨逊夫人》7 篇文章中的 600 个大句，汉语约 2.88 万字。测试结果如下：

机器/人工	无头迭句	花园幽径句A	花园幽径句B	花园幽径句C	普通迭句	其他	总计	准确率②（%）
无头迭句	41		1	2	2		46	89.1
花园幽径句A		7					7	100.0
花园幽径句B			17		1	6	24	70.8
花园幽径句C	11			8	2	3	24	33.3
普通迭句	1		1	1	126	37	166	75.9
其他	7	1		6	38	281	333	84.4
总计	60	8	19	17	169	327	600	80.0
召回率①（%）	68.3	87.5	89.5	47.1	74.6	85.9	80.0	

注①、召回率=机器判定与人工判定一致的文本数/人工判定该类文本的总数；

注②、准确率=机器判定与人工判定一致的文本数/机器判定该类文本的总数；

注③、表中空白处代表该统计结果为0。

首先，对于字词层面能够激活识别的句式，包括普通迭句、无头迭句、花园幽径句A和花园幽径句B，只需依赖词语知识库中的“概念类别”信息即可使识别有较高的召回率和准确率。

其次，对于需要依赖广义作用句和广义效应句知识来辨识的句式，即花园幽径句C，本文以“句类代码”中不同标号作为广义作用词和广义效应词的识别标志，在识别上却造成了与无头迭句的混淆，自身的召回率和准确率不高，同时还影响到无头迭句的识别结果。究其原因，存在一部分词语的“句类代码”同时含有两种标号，故造成识别混乱。事实上，花园幽径句C的识别应结合“HNC符号”中的相关知识来识别，在此限于篇幅未能予以讨论。

最后，技术指标的进一步优化有待于：所用词语知识库的完善（尚存在一些影响识别的字词未予以收录）；实现EK复合构成的正确辨识（本次识别算法仅实现EK块的最简识别）。

4.2 实例翻译结果

对引言中的例子（是为花园幽径句C）采用所提出的转换规则进行翻译如下：

We should adopt a scientific approach and make bold explorations in order that we think and act more in line with the objective reality, China's conditions in the primary stage of socialism, and the development of the times.

可以看出，经过句式转换处理后，翻译结果结构正确且可读性强，因此，若在基于规则的翻译系统中考虑这些特殊句式，相信能够适当提高该系统的翻译准确率。

5 结语

此次探讨的句式都是从字词层面即能激活辨识的类型，包括迭句和三类花园幽径句，但还不够全面，如花园幽径句就未囊括所有情况。

在句式识别算法的设计和实现方面，仍有改进之处，表现在：识别程序中多次进行了从简处理，更多的是考虑简单常见的情况；仅处理了一小部分的句式，而现实中已经发现并提出的句式不在少数。因此，在算法中考虑更多的句式，程序上实现常规判断，以提高句式识别系统的适用性，这是下一阶段可尝试的工作。

参考文献

[1] 冯志伟. 花园幽径句的自动分析算法.当代语言学,2003(4):339-349.

[2] 池毓焕. 汉语动词形态困扰的分析与处理.北京:中国科学院声学研究所博士学位论文,2005.

[3] 黄曾阳. HNC（概念层次网络）理论.北京:清华大学出版社,1998.

[4] 黄曾阳. 语言概念空间的基本定理和数学物理表示式.北京:海洋出版社,2004.

[5] 池毓焕,李颖. 面向汉英机器翻译的大句范式初探. 孙茂松,陈群秀主编:中国计算语言学研究前沿进展(2007-2009),北京: 清华大学出版社,2009,395-400.

[6] 李颖,王侃,池毓焕. 面向汉英机器翻译的语义块构成变换.北京:科学出版社,2009.

[7] 张克亮. 汉英机器翻译中是否判断句的句类转换. 黄河燕主编:机器翻译研究进展—2002 年全国机器翻译研讨会论文集,北京:电子工业出版社,2002,172-183.

[8] 张克亮. 面向机器翻译的汉英句类及句式转换,开封:河南大学出版社,2007.

[9] 李颖,池毓焕. 基于机器翻译的原型句蜕及其包装研究.装甲兵工程学院学报,2003,17(3):7-13.

[10] 李颖,池毓焕. 句类的扩展表示. 朱小健,张全,陈小盟主编:中文信息处理的探索与实践,北京: 北京师范大学出版社,2006,371-378.