

面向汉英机器翻译的大句范式初探*

池毓焕¹ 李颖²

1. 中国科学院声学研究所, 北京 100190; 2. 装甲兵工程学院信息工程系, 北京 100072

E-mail: chiyuhuan@hotmail.com; lypublic@hotmail.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。本文初步探讨了汉英机器翻译面临的几个常用大句范式, 描述其辨识特征, 并提出转换规则, 以期对现有基于规则的汉英机器翻译系统有所助益。

关键词: 大句范式, 格式与样式, HNC 机器翻译引擎, 汉英机器翻译

A Basic Research on the Paradigm of Major Sentence in Chinese-English Machine Translation

Chi Yuhuan¹ Li Ying²

1. Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China;

2. Department of Information Engineering, Academy of Armored Force Engineering, Beijing 100072, China

E-mail: chiyuhuan@hotmail.com; lypublic@hotmail.com

Abstract: In the viewpoint of Major Sentence, there should be some certain modes how to organize its minor sentences that are named as Paradigm of Major Sentence. Furthermore, there are some remarkable differences of using those paradigms of major sentence between different languages. So they should be transformed while translating. In this paper, some frequently-used paradigms of major sentence are discussed while the characteristic how to distinguishes them and transforming rules put forward in order to help those rule-based Chinese-English machine translation systems.

Keywords: Paradigm of Major Sentence, Format and Pattern, HNCMT Engine, Chinese-English Machine Translation.

1 引言

现阶段的机器翻译应超越逗号而以句号为基本处理单位。在句号的视野内会呈现出某些特定的结构模式, 而模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。

我们把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^[1]。约 65% 的语串成句。如果这些语句仅是语段的构件, 则称之为小句; 相应的, 由若干小句构成的语段称作大句。

此前, HNC 理论^[2-4]已为基本句、单句或小句引入了格式和样式的概念。语句格式指广义作用句各主语义块位置的不同排列组合; 语句样式指广义效应句各语义块位置的可能排列组合。广义作用句主块位序的变动需要借助(但不是必须借助)语义块标记符号的配合, 而广义效应句则不需要借助(一定不借助)语义块标记符号的配合。

最近, 为与此相区别, 黄曾阳先生特地引入了大句范式的概念, 以表征大句的句式结构。例如“是{!31EIJ}的。”是汉语相当常见的大句范式之一, 而英语根本不存在这种大句范式, 翻译时就要进行相应的大句范式转换。

具有形式特征的大句接近于语言学家的句群^[5-7], 相关理论探索无疑颇具启迪意义。而“以章句为始基”正是 HNC 理论继承于训诂学的基本理念之一。面向自然语言理解的技术侧面, 力

* 本文承国家科技支撑计划(NO.2007BAH05B02-05); 国家重点基础研究发展规划(973)(No.2004CB318104); 中科院声学所知识创新工程项目(No.0654091431); 中国科学院声学研究所“所长择优基金”(No.GS13SJJ04); 中国科学院青年人才领域前沿项目(No.O754021432)的资助。

求对语言现象进行形式化描述, HNC 与传统语言学的面向和定位之别泾渭分明。

本文初步探讨四个汉英机器翻译中常见的大句范式(因尚未分别命名, 以下权以编号称之), 均按特征描述、说明、转换规则、例句、对现有基于规则的汉英机器翻译系统的测试等部分展开, 以图对现有系统有所助益。

2 范式一

自然语言描述: 以“是”为 Eg(Eigen global, 近似于顶层谓语), 以“的”结束, 两者之间的 DC 是一个或多个原型句蜕; 多个原型句蜕之间有逗号, 各原型句蜕共享位于大句之尾的“的”。

半形式化描述: 范式 1-1——“是{!31EIJ}的。”; 范式 1-2——“是{!32EIJ}的。”。‘|’表示此前的构成单位可重复, 其它除了自然语言符号便是 HNC 语句标注符号, 后者可参阅文献^[8]。

说明: “是-的”结构存在若干变种, 如若干个“是-的”结构并列共享 GBK₁(Generalized oBject chunK 1st, 近似于主语), 单个“是-的”结构拖带若干个原型句蜕等。

转换规则: 去掉“是”和“的”, 把原来的 El(Eigen local, 近似于从句谓语)提升为 Eg; 对于范式 1-2, 还要进行格式变换, 即!32=>!021。是否判断句转换的已有研究可参阅文献^[9-10]。

例 1. 它||是||{反对|帝国主义压迫}, {主张|中华民族的尊严和独立}的。(f14\2)

It || opposes || imperialist oppression + and upholds || the dignity and independence of the Chinese nation.

例 2. 在当时, ~||这种所谓新学的思想, ||有||\{同中国封建思想|作斗争}的革命作用/, +是||{替旧时期的中国资产阶级民主革命|服务}的。(f14\2)

At the time, ~|| the ideology of the new learning || played a revolutionary role in || {fighting | the Chinese feudal ideology}, ++and it || served || the bourgeois-democratic revolution of the old period.

例 3. 电影《红高粱》||是||{张艺谋|导演}, {姜文、巩俐|主演}的。(f14\2)

The film "Red Sorghum" || was directed || by Zhang Yimou + and starred || by Jiang Wen and Gong Li.

例 1 为“是{!31EIJ}₂的。”结构, 其中的下标表示原型句蜕并行出现的次数, 下同。例 2 虽然仍是!31 格式, 即省略 GBK₁, 但本身却是!111 格式, 因此要进行规范格式变换。例 3 为“是{!32EIJ}₂的。”结构, “导演”动名兼性、以名为主, 需要与“主演”协调才能识别这一结构。

经测试, 现有系统基本上认识大句范式一, 单项处理正确率接近 70%, 错误主要发生在范式 1-2, 即没有认出!32 格式或认出了也不知道向英语被动式转换。

3 范式二

自然语言描述: 全部是!310 格式的小句, 或一系列!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”)起头的小句。这两种基本形态还各有两种瘦身形态, 其一是以一系列动宾结构或主谓结构打头阵; 其二是以一系列类似标题语的东西打头阵。

半形式化描述: 范式 2-1——“{!310EIJ}!310EgJ。”; 范式 2-2——“{!310EIJ}({f84})EgJ。”

说明: 范式 2-1 实质上就是花园幽径句, 可参见文献^[11]。因形式上与无头迭句相同, 问题转化为如何从无头迭句中辨识出花园幽径句。试比较:

例 4. 坚持实施||可持续发展战略, +正确处理||经济发展同人口、资源、环境的关系, +改善||生态环境+和美化||生活环境, +改善||公共设施和社会福利设施。

We || will adhere to || the strategy of sustainable development + and correctly handle || the relations between economic development on the one hand and population, resources and the environment on the other, + improve || the eco-system + and beautify || the living environment, + and improve || public and social welfare facilities.

例 5. {清理|古代文化的发展过程}, {剔除|其封建性的糟粕}, {吸收|其民主性的精华}, ||是||\{发展|民族新文化}{提高|民族自信心}的必要条件/; 但是[[决不能无批判地兼收并蓄]].

{To study | the development of this old culture}, {to reject | its feudal dross} and { assimilate | its democratic essence}||

is \parallel a necessary condition for { developing | our new national culture } and { increasing | our national self-confidence }, ++ but we \parallel should never swallow \parallel anything and everything [+uncritically +].

例 6. {敏锐地把握 \parallel <我国社会生产力|的发展>趋势和要求}, +{坚持 \parallel {以经济建设|为中心}}, +{不断促进 \parallel <先进生产力|的发展>}, (这 f84)是 \parallel {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/。

We \parallel must have a profound understanding of \parallel the development trend and requirements of our country's social productive forces, + focus on \parallel economic development + and promote \parallel the constant development of the advanced productive forces. +++[Only by doing so] \parallel [can] we \parallel really ensure \parallel that {our Party | always stand in | the forefront of the times} and { maintain | its advanced nature }.

例 4 的四个小句（实质上是五个小句）全部省略 GBK₁，缺省的主语是“我们”或“我党”，在讲话或行文中不言自明。后续小句共享前一小句的 GBK₁，称作迭句；若被看齐的小句也省略 GBK₁，则称无头迭句，可记作 1310[EgJ]_m。英语不允许无头迭句，因此要求“补头”，即加上代词如例 4 和例 6 所示，或向施事不必明说的被动式 I02 变换。例 5 的前三个小句是 {1310EJ}₃，共同构成第四小句的 GBK₁，故为系列花园幽径句。例 6 在‘是’前加了个 f84 ‘这’。

从大句的视野看，并列小句具有句类和格式的协调性。若系列小句都属于广义作用句，格式相同，接着出现一个广义效应句——特别是基本判断句——加以总结，则此前的系列小句一同降为原型句蜕；或者系列小句描述广义效应的各侧面，并且样式相同，接着来一个基本判断句，花园幽径句也。故辨识系列花园幽径句的策略为检验各小句的句类，基本判断句的出现则加强之。

转换规则：按常规的原型句蜕变换规则^[12]处理打头的若干并列小句。因为由若干小句构成的 GBK₁ 偏于复杂，通常向英语的特定句式^[10]转换，即引导词 it 作形式主语先翻主句，把系列不定式短语或现在分词短语后置。也可进行‘是’字句的 DB-DC 换位处理，以避免头重脚轻的现象。

例 7. {始终保持 \parallel 同人民群众的血肉联系}, \parallel 是 \parallel {我们党|战胜|各种困难和风险}、+{不断取得|事业成功}的根本保证/。

{ Maintaining | the flesh-and-blood ties with the masses of the people } \parallel is \parallel the fundamental guarantee { for our Party | to overcome | difficulties and risks } + and { make continuous success in | our cause } /.

例 8 {发展 \parallel 党内民主}, +{充分发挥 \parallel 广大党员和各级党组织的积极性主动性创造性}, \parallel 是 \parallel {党的事业|兴旺发达}的重要保证/。

[{It | is | imperative}] to promote \parallel inner-Party democracy + and give full play to \parallel the enthusiasm, initiative and creativity of the Party members and Party organizations at all levels. +++ This \parallel is \parallel an important guarantee for the success of the Party cause.

例 9. 在新的世纪, ~ \parallel {继续推进|现代化建设}, +{完成|祖国统一大业}, +{维护|世界和平} + 与 {促进|共同发展}, \parallel 是 \parallel <我们党|肩负|的重大历史任务>。

In the new century, ~ \parallel the great historical tasks for our Party \parallel are \parallel { to continue | the modernization drive }, { accomplish | the great cause of the reunification of our motherland }, { safeguard | world peace } and { promote | common development }.

例 7 的 GBK₁ 原型句蜕向现在分词短语变换。例 8 向特定句式转换，而主句独立成句，加 this 以指代。例 9 属于‘是’字句的 DB-DC 换位。

经测试，现有系统对大句范式二仅有部分认识，乃与‘是’字句一同处理，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；另一个系统则知道在‘是’前加 it，但不知道为此前的系列小句“安头”并单独成大句。

4 范式三

自然语言描述：同一个动词重复使用而形成一串迭句。这个重复动词常常是“是”，也可以是“有”、“要”以及“判断概念群 8”的某些词语（如“主张”）。故本范式可按重复动词的类型

和宾语的类型分成若干子类。

范式 3-1 半形式化描述：“ETJ+(!310ETJ)”，其中 ET 表示以同一动词作为 Eg。这是范式三的基本类型，意即：除非符合其他子类，符合范式三的都按范式 3-1 处理。

说明：可能受骈文的影响，汉语喜欢重复谓语，构成排比句，似乎因此而加强了气势。但英语的谓语重复相当罕见，因此多数要进行汉英“小句多—转换”，把重复的谓语合并，形式上表现为汉语的多个迭句转换成英语的单句。本范式同样适用于 EIJ 中重复谓语的连迭现象。

范式 3-1 转换规则：保留第一小句 ETJ，并且新的 $GBK_2' = \Sigma(GBK_{2m})$ 。

例 10. 我们||[要]在<党的基本理论、基本路线、基本纲领的指引>下||，继续坚持和完善||{公有制为|主体}、{多种所有制经济共同发展}的基本经济制度|，+坚持和完善||社会主义市场经济体制|，+坚持和完善||<按劳分配>为|主体的多种分配方式>|，+坚持和完善||对外开放|；

We ||[-must-], under the guidance of the Party's basic theories, basic line and basic program, ~|| stick to and improve || the basic economic system with { the public ownership | playing a dominant role} and { all forms of ownerships | developing side by side}; the socialist market system; the diversified forms of distribution with the distribution according to work as the main modality; the opening-up program;

例 11. 它||是||{反对|一切封建思想和迷信思想}，{主张|{实事~|求|是}}，{主张|客观真理}，{主张|{*理论和实际|一致}}的。(f14\2)

[-Opposed as-] it || is to || all feudal and superstitious ideas, ++it || stands for || {seeking | truth | from facts}, [+for+] objective truth and [+for+] the unity of theory and practice.

例 11 总体为“是{!31EIJ}₄的。”结构，其中以“主张”为 EK 的{!31EIJ}₃应用了范式三。

范式 3-2 半形式化描述：“jDJ+jDJ(DC_m∈MLC)”，为两小句结构，其中 Eg 同为“是”，DC_m∈MLC 表明它的宾语属于多元逻辑组合(Multiple Logic Combination)，既不是单词，也不是句蜕。

说明：在英语中 be 作为 Eg 进行重复看来并非罕见，故列为特殊子类。

范式 3-2 转换规则：保留第一小句 jDJ，并改逗号为句号；为第二小句“安头”It 或 They。

例 12. 生产力||是||最活跃最革命的因素，+是||社会发展的最终决定力量。

The productive forces || are || the most dynamic and revolutionary factor. +++They || are || the ultimate decisive force of social development.

例 13. 马克思主义||是||我们立党立国的根本指导思想，+是||全国各族人民团结奋斗的共同理论基础。

Marxism || is || the fundamental guiding principle for the consolidation of the Party and the development of the country. +++ It || also constitutes || the common theoretical foundation of the concerted efforts of the people of all ethnic groups.

范式 3-3-1 半形式化描述：“jDJ+(!310jDJ)(DC_m={!310EIJ})”，其中 jDJ 的 Eg 同为“就是、不是、没有”等，DC_m={!310EIJ}表明它的宾语为系列!31 基本格式的原型句蜕。

说明：当汉语重复谓语的迭句用于表示强调时，英语也可以出现重复谓语的句式表示强调，即进行“迭句-列句转换+大句-句群转换”。这时英语的列句一定是 jDJ/jD1J 句类，其 DB 一定是代词。如果汉语的 DC 为原型句蜕，则要进行“原型句蜕-逻辑组合变换”^[13]；如果汉语的 DC 原型句蜕加了包装^[14]，则英语的 DC 可能为从句。

范式 3-3-1 转换规则：保留第一小句 jDJ；为后续小句“安头”；进行原蜕-逻辑组合变换。

例 14. {对于这些人，|如果不加以惩罚}~|，我们||就是||{犯错误}，+就是||{纵容汉奸}，+就是||{不忠实于|民族抗战}，+就是||{不忠实于|祖国}，+就是||{纵容|坏蛋|来破裂|统一战线}，+就是||{违背了|党的政策}。

\Failure { to punish | them } || would be || a mistake; +++ it || would be || an encouragement to the collaborators and traitors, ++ it || would be || disloyalty to the national resistance and to our motherland, and an invitation to the scoundrels to disrupt the united front. +++ It || would be || a violation of the policy of our party.

范式 3-3-2 半形式化描述：“jDJ+(!310jDJ)(DC_m={!310EIJ})”，其中 jDJ 的 Eg 同为“是、不是”，DC_m={!310EIJ}表明它的宾语为系列!31 基本格式的原型句蜕。

说明：范式 3-1 中系列原型句蜕汉英变换常存在“原型句蜕-ing 短语变换”或“原型句蜕-to 短语变换”的选择困难，因为二者的主要区别在于 ing 短语表过程或进行时态，to 短语有表目的或将来时态，而机器难以把握这一点。一旦与‘是’字句转换联系起来，二者的选择就明确了：若‘是’字句进行句类零转换，而“be+现在分词”肯定用于表示进行时态，则只好选择“be to”结构；若进行句类转换，则选用 ing 短语还是 to 短语取决于转换后的 Eg 动词的语习。

范式 3-3-2 转换规则：‘是’字句句类转换，并且新的 $GBK_2' = \Sigma(GBK_{2m})$ ，其中 GBK_{2m} 进行“原型句蜕-ing 短语变换”或“原型句蜕-to 短语变换”；‘是’字句零转换，保留第一小句 jDJ，并且新的 $DC' = \Sigma(DC_m)$ ，其中 DC_m 要进行“原型句蜕-to 短语变换”。

例 15. 但是这种尊重，||是||{给历史以一定的科学的地位}，+是||{尊重}历史的辩证法的发展}，+而不是||{颂古非今}，+不是||{赞扬任何封建的毒素}。

However, respect for history || means || {giving | it | its proper place as a science}, +{respecting | its dialectical development}, +and {not eulogizing | the past | -at the expense of the present } +or {praising | every drop of feudal poison}.

例 16 在当代中国，~||{发展先进文化}，||就是||{发展有中国特色社会主义的文化}，+就是||{建设社会主义精神文明}。

In contemporary China, ~|| {to develop | advanced culture} || is || {to develop | culture with distinct Chinese characteristics} +and {to build | socialist spiritual civilization}.

范式 3-3-3 半形式化描述：“jDP21*21J+(!310jDP21*21J)(PBC2_m={!310EIJ})”，其中 jDP21*21J 表明 Eg 同为“是为了”或“不是为了”，即汉语的 EK 为复合结构。

说明：同范式 3-3-2，单列一子类主要为了突出复合结构“是为了”可视为一个重复谓语，肯定要转换为英语的 be+to 结构。

范式 3-3-3 转换规则：保留第一小句 EIJ，并且新的 $GBK_2' = \Sigma(GBK_{2m})$ 。其中 GBK_{2m} 要进行“原型句蜕-to 短语变换”。

例 17. 但是这种{给投降派和反共顽固派以打击}的政策，||全是为了||{坚持抗日}，+全是为了||{保护抗日统一战线}。

However, the sole purpose of \ the policy of {dealing blows to | the capitulators and the anti-Communist die-hards}/ || is ||{to keep up| the resistance to Japan} +and {safeguard | the anti-Japanese united front}.

经测试，现有系统在大句范式三上几乎全军覆没，其中最优秀的系统能进行范式 3-1 变换，但没能认出作为特例的其他几个子范式。

5 范式四

自然语言描述：汉语系列迭句形态的 Eg 表现为成对的反义动词如“是……不是……”、“不是……而是……”、“要……不要……”、“支持……反对”等。这种结构可称作反义对仗句。

半形式化描述：“ $GBK_1(!310ETJ+E^TJ)$ ”，其中 ET 表示以集合 T 中的同一动词作为 Eg， E^T 表示与 ET 对应的反义动词作为另一 Eg，并且 $T = \{(是, 不是); (有, 没有); (要, 不要); (主张, 反对); \dots\}$ 。

说明：汉语的‘是、有、要’等分实用（直接充当 EK）和虚用（充当 EK 一部分的 QE 等）。

反义对仗句关键在于虚实用的辨识^[1]和其他动词角色的确认。一旦认出属于实用，其 GBK_2 构成的变换规则就相对确定下来。同时 j1 类概念还适用范式 3-1 的避免重复转换。

转换规则：保留第一小句 ETJ，新的 $GBK_2' = \Sigma(GBK_{2m})$ ，其中 GBK_{2m} 之间用 but, instead of 等联结。如果反义对仗句的标记词实用，若任一小句未跟随动词，则存在于其他小句的原型句蜕进行“原型句蜕-逻辑组合变换”；若未被名词化，则进行“原型句蜕-to 短语变换”。

例 18. {要相互尊重与平等互利}，{不要霸权主义和强权政治}；{要对话与合作}，{不要对抗与冲突}，已成为||越来越多国家的共识。

[It] has become || the common understanding of a growing number of countries || {to embrace | mutual respect, equality and mutual benefit } +and {reject | hegemonism and power politics}, +{ to pursue | dialogue and cooperation } +and {avoid | confrontation and conflict}.

例 19. 国民党的反共顽固派||强调||统一, ++但是[他们的所谓“统一”], ||乃是||假统一, +不是||真统一; +乃是||不合理的统一, +不是||合理的统一; +乃是||形式上的统一, +不是||实际的统一。

The anti-Communist die-hards within the Kuomintang || emphasize || unification, ++but their so-called unification || is || not genuine but a sham, not a rational but an irrational unification, not a unification in substance but in form.

例 20. 他们||反对||“城市中心论”, +主张||{向敌人力量薄弱的广大农村|发展}; +反对||军事冒险主义, +主张||诱敌深入; +反对||{用\{削弱\地方武装\}的方法来\{来扩大\主力红军\}, +主张||{两种武装力量|都要发展};

They || opposed || the theory of "making cities the centre of the Chinese revolution" + and advocated || {building | strength |~ in the vast rural areas, where the enemy's forces were relatively weak}. +++They || rejected || military adventuresome ||~ in favor of {luring | the enemy [+in deep+]}. +++They || were against || {expanding | the Red Army's main forces |~ at the expense of local armed forces} +and urged || that {both | be expanded simultaneously}.

例 18 靠前面系列广义作用句(‘要’实用)加双对象效应句(‘成为’)来辨识出范式二, 其中复杂的 GBK₁ 又适用范式四, 而全句向特定句式转换。例 20 存在双层对仗: 分号表示的对仗和分号内“反对”与“主张”的对仗。

经测试, 现有系统对“是……不是……”或“不是……而是……”等句式还是有所认识的, 即看作虚化的连词, 而对实用现象的辨识和变换又几乎全军尽没。

6 结语

上述大句范式都是从字词层面即能激活辨识的范式类型, 实际上还有其他类型未能论及。

大句范式的覆盖率, 意味着对提高译准率的作用范围, 是尚待我们进行研究的基础性工作, 这也是本文题为“初探”的原因所在。但能像本文这样准确描述的大句范式肯定是有限的。

汉英翻译之范式转换的必要性植根于汉英两种语言存在的本质差异。与之相关的汉英差异, 主要是“汉语比较重视并擅长默认省略方式, 而英语比较重视并擅长指代替换方式”和“汉语偏好精练的小句, 无非限短语, 无从句, 而英语偏好冗长的句子, 非限短语和从句满天飞”。

参 考 文 献

- [1] 池毓焕. 汉语动词形态困扰的分析与处理[D].北京:中国科学院声学研究所博士学位论文,2005.
- [2] 黄曾阳. HNC(概念层次网络)理论[M].北京:清华大学出版社,1998.
- [3] 黄曾阳. 语言概念空间的基本定理和数学物理表示式[M].北京:海洋出版社,2004.
- [4] 苗传江. HNC 理论导论[M].北京:清华大学出版社,2005.
- [5] 曹政. 句群初探[M].杭州:浙江教育出版社,1984.
- [6] 王缙. 复句·句群·篇章[M].西安:陕西人民出版社,1985.
- [7] 吴为章,田小琳. 汉语句群[M].北京:商务印书馆,2000.
- [8] 李颖,王侃,池毓焕. 面向汉英机器翻译的语义块构成变换[M].北京:科学出版社,2009.
- [9] 张克亮. 汉英机器翻译中是否判断句的句类转换[A]. 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[C],北京:电子工业出版社,2002,172-183.
- [10] 张克亮. 面向机器翻译的汉英句类及句式转换[M],开封:河南大学出版社,2007.
- [11] 冯志伟. 花园幽径句的自动分析算法[J],当代语言学,2003(4):339-349.
- [12] 李颖,池毓焕. 基于机器翻译的原型句蜕及其包装研究[J],装甲兵工程学院学报,2003,17(3):7-13.
- [13] 池毓焕,李颖. 汉语原型句蜕向英语多元逻辑组合的变换[J],装甲兵工程学院学报,2004,18(3):74-78.
- [14] 李颖. 面向汉英机器翻译的包装句蜕处理[J],计算机应用,2004,24(6):137-141.