文章编号：1003-0077（2011）00-0000-00

汉英机器翻译中格式转换研究*

刘智颖[1,3]，郭艳波[2]，晋耀红[1,3]

（1.北京师范大学中文信息处理研究所，北京 100875；

2. 盘古文化传播有限公司，北京 100162；

3. 中国专利信息中心-北京师范大学机器翻译联合实验室， 北京 100875）

摘要：格式在HNC理论中是指广义作用句各主语块位置的不同排列组合方式。由于主语块的排列方式在汉英两种语言中表达的差异，汉语句子翻译到英语时常常发生格式转换。格式转换是HNC机器翻译理论的一个重要内容，是机器翻译理论实践的基础和前提。本文以HNC机器翻译理论为指导，以真实文本的专利文献汉英句对为分析对象，研究专利机器翻译中汉英两种语言之间广义作用句的格式转换规律，制定了排除规则、识别规则和转换规则，并对部分规则进行了人工评测，结果表明准确率能达到85%左右。

关键词：格式转换；广义作用句；机器翻译

中图分类号：TP391　　**文献标识码：**A

The Format Conversion in Chinese-English Machine Translation

LIU Zhiying[1,3], GUO Yanbo[2] ,JIN Yaohong[1,3]

(1.Institute of Chinese Information Processing, Beijing Normal University, Beijing, 100875,China;

2. Pangu Culture Media Company, Beijing, 100162, China;

3. CPIC-BNU Joint Laboratory of Machine Translation, Beijing Normal University, Beijing, 100875, China)

Abstract: The format in Hierarchical Network of Concepts (HNC) theory refers to the different arrangement method of the main chunk in the general action sentence. The format conversion always occurs in the Chinese-English translation for the difference of the main chunk arrangements in two languages. Based on the HNC machine translation theory, this paper analyzed the patent documents Chinese-English sentence pairs, studied the format conversion laws in the general action sentences, made the exclusion rules, recognition rules and conversion rules, evaluated the effects of part rules. Our experiments show that we can obtain a translation precision of about 85%.

Key words: format conversion; general action sentence; machine translation

1 引言

格式转换在HNC理论[1]中是指广义作用句各主语块位置的不同排列组合方式。由于主语块的排列方式在汉英两种语言中表达的差异，汉语句子翻译到英语时常常发生格式转换。格式转换是HNC机器翻译理论的一个重要内容，是机器翻译理论实践的基础和前提。

HNC理论以概念联想脉络为主线，建立了自然语言的计算机理解处理模式，该理论的一个重要应用之一就是研究和开发汉英机器翻译系统。HNC机器翻译是基于规则的机器翻译系统，分为源语言分析、过渡处理、目标语生成三个处理阶段。过渡处理包括六个环节，即句类转换、句式转换、主辅语块变换、语块构成变换、辅块排序调整和小句排序调整[2]。其中，句式转换包括格式转换和样式转换。格式转换存在于广义作用句中，而样式转换存在于广义效应句中。据统计，汉英机器翻译中，需要进行格式转换的句子占20%[3]。可见，格式转换是机器翻译的一个重要内容。

本文以HNC机器翻译理论为指导，以真实文本的专利文献汉英句对为分析对象，从HNC

* **收稿日期：2013-06-10　定稿日期：2013-07-15**

基金项目：国家高技术研究发展计划（863）（2012AA011104）；中央高校基本科研业务费专项资金

作者简介：刘智颖（1975——），女，博士，主要研究方向为中文信息处理；郭艳波（1987——），女，硕士，主要研究方向为中文信息处理；晋耀红（1973——），男，教授，主要研究方向为信号与信息处理。

角度研究专利机器翻译中汉英两种语言之间广义作用句的格式转换规则，包括排除规则、识别规则和转换规则。经过测试，语义翻译引擎对格式转换的处理取得良好的效果，对于全局的格式转换处理的准确率能够达到85%左右。

2 相关工作

在世界上的语言中，按句子语序可分为三种类型：主动宾(SVO)、动主宾(VSO)、主宾动(SOV)。英语的语序多为主动宾(SVO)，现代汉语在语序类型上属于SVO型语言，语法上的一般规则是：句子成分一般按照“主语——谓语——宾语”的顺序排列。而汉语的语序很大程度上取决于句子的意义，因而主语与动词的次序较为灵活[4]。

格式转换又叫调序，即根据需要调整句子的语序[5]。调序在统计机器翻译中是很重要的一个环节，调序方法主要有两类：采用概率统计方法和采用模版方法。各种调序模型及对调序模型的融合研究逐渐成为机器翻译研究的热点[6]。

HNC理论对格式问题也做过相应的研究。针对某种特定句类，曾经研究过汉英翻译中一般转移句的格式转换，总结了一般转移句格式转换的规律[7]，块扩句式转换问题[8]。针对英机器翻译中的句式转换，研究了汉英两种语言在句式表达方面的异同，描述了汉英句式转换的一般规律[9]。此外，还就汉英机器翻译的格式自转换进行了研究[10]。不过，这些研究也仅停留在理论研究和构想阶段，对语言现象分析是理论层面的，制定的形式化规则没有得到实验验证，而且在分类上可以更细。

本文在以上研究的基础上，对汉英专利机器翻译的格式转换进行更深入、更全面、更具体的研究，所制定的转换规则直接服务于汉英专利机器翻译语义引擎，并可以在语义引擎中直接检验规则的有效性，从而实现对规则的实时调试与修改。

3 汉英格式问题

格式，又叫语句格式，是指句子中主语块的排列顺序[11]。句类表示式说明了一个句类由几个什么样的主语块构成，而这些主语块在不同的句子中可能顺序不同，这就是语句格式的不同。

在HNC理论中，不考虑语块的省略，语句格式有三种类型：

（1）基本格式（!0）：对于三主块句，句子的格式是“GBK1+EK+GBK2”。也就是SVO的格式。

（2）规范格式（!1）：对于三主块句，句子的格式是“GBK1+^GBK2+EK”或“GBK2+^GBK1+EK”。也就是SOV或OSV格式，广义对象语块（S和O）相邻且相邻语块之间存在语块标记。

（3）违例格式（!2）：不同于规范格式，广义对象语块相邻且相邻语块之间不存在语块标记。

句类分为广义作用句和广义效应句两大类。只有广义作用句才具有格式信息。

对于广义作用句而言，汉语既允许使用基本格式和违例格式，也允许使用规范格式，对某些句类甚至偏好规范格式，如：主动反应句；而英语只允许使用基本格式或违例格式，不允许使用规范格式，因为形成规范格式所必需的语法工具（即HNC所定义的语言逻辑10概念）英语是残缺不全的，而汉语是完备的。

汉语中，概念林10辖属4株概念树，分别作为不同类型语块的标识符。

L0概念分类	**功能说明**	**例词**	**例句**
L00	特征语块（E）标记	了、着、过、而、所	本发明公开**了**处理浸过水的种子的方法。
L01	作用者（A）语块标记	由、被、给	其标准化工作**由**第三代合作项目（3GPP）组织完成。
L02	对象语块（B）标记	把、对、向、为	加密单元不**对**存储数据进行解密。
L03	内容语块（C）标记	把、将	印刷装置148把标签剂C混合物印刷到衬底114上。

表 1：概念层次网络理论中的概念林 L0 分类

英语的广义作用句不存在规范格式。当汉语句子的规范格式翻译到英语时，必然发生格式转换。如汉语句子“播放器对该内容进行解扰。（The player descrambles the content.）”，采用的是规范格式“GBK1+^GBK2+EK”，英语采用基本格式“GBK1+EK+GBK2”。

由于规范格式存在明显的语块边界标识符，所以本文着重研究汉语广义作用句的规范格式向英语的转换问题。

本文的研究单位是以逗号或句号划分成的单句或小句。格式转换既可能发生在单句和小句中，也可能发生在单句或小句内部的语块中。本文关注前者，即发生在单句中的格式转换。研究的前提是小句已经切分，EG（特征语块）、ABK（辅块）、LB（句间逻辑说明符）已经识别出来。

4 语料分析与标注

本文的研究对象是汉英专利机器翻译广义作用句的格式转换，语料使用中国专利信息中心的检索系统根据 10 概念（将、把、对、向等）检索出来的 1 万句汉英句。

本文对语料的标注是多维度的，包括格式转换的现象、依据、结果和规则。现象指源语言中的语言逻辑概念（10）和特征语块（E）。依据是影响格式转换的因素，包括句类因素、是否有 JK1、是否有联结词、是否发生句类转换等。结果描述汉语句子翻译到英语句子后，是主动形式还是被动形式，以及翻译前后源语言和目标语的格式变化。规则部分用较为简练的符号标注了格式转换的条件及结果，“=>”左边是条件，右边是结果。

中文语料	参考英译	现象	依据	结果	规则
		L0 E	句类	格式代码	
第二通信模块将第二格式的第二表示数据发送到计算机系统。	The second communications module transmits the second indicating data in a second format to the computer system.	将 发送到	T0	!114	将 &l03+T0 =>!114
这些计数器对这些数据输入 / 输出装置发出的总线分配请求数进行计数。	These counters count the number of bus allocation request signals issued from these data input/output devices.	对 计数	X	!11	对&l02+X =>!11

表 2：格式转换语料的多维标注

对语料进行标注分析，总结规则后，要对规则进行形式化，便于计算机识别和处理。为此设立了一套规则符号，包括特征集、位置标记、操作函数、属性集等。定义好规则符号后，即可对规则进行形式化表示。例如：

“这些计数器||**对**这些数据输入装置发出的总线分配请求数||进行计数。”			
位置： (-1)	(0)	(1)	(2)
语块： GBK	*l0*	GBK	EG
语义属性：	LEVEL=1		V_COMP
These counters ‖count‖ the number of bus allocation request signals issued from these data input/output devices.			
位置： (-1)	(2)	(1)	
语块： GBK	EG	GBK	
规则：(-1)LC_CHK[GBK]+(0)CHN[对]&LC_CC[*l0*]&LC_LEVEL[1]+(1)LC_CHK[GBK]+(2)LC_CHK[EG]=>(-1)+DEL_NODE(0)+(2)+(1)$			
解释：这条规则的意思是，将相对位置(1)表示的语块移到(2)后面，并删除(0)位置的语言逻辑概念词“对”。			

表 3：格式转换语料规则表示

5 格式转换规则

汉英专利机器翻译格式转换规则研究，包括研究其排除规则、识别规则和转换规则。

排除规则主要是排除与 10 概念兼类的其他概念，充当 10 概念的通常是汉字“把、将、对、向”等，但这些词不仅充当 10 概念，还充当动态概念、基本概念等。所以首先要对这些不属于 10 概念的情况进行排除，识别出 10 概念。

识别规则主要是识别 10 的层次，单句中 10 的层次记为 1，小句中 10 的层次记为 2，不同层次格式转换的规律不同，所以要对 10 的层次进行识别。

最后制定格式转换规则。

不管是排除规则、识别规则还是转换规则，都具有一定的优先顺序。首先，排除规则优先于识别规则和转换规则；其次，所有规则都以(0)号节点（通常为 10 概念）为切入点，先向前匹配，再向后匹配。

5．1 排除规则

充当 l0 概念的词都是常用词，几乎都具有兼类现象，所以要先进行处理，排除含 l0 概念的词但不属于格式转换的情况。可以利用的信息有：

- EG 信息

l0 概念的词大体对应于介词，大多具有动态概念属性，下面这条规则可统一排除这种兼类情况。

(0)CHN[把,将,对,向,由,给,比,与]+(f){!LC_CHK[EG]}=>!LC_SELECT(0,LC_CC,l0)$

此条规则的含义是：当“把，将，对，向，由，给，比，与”后面找不到特征语块（EG）时，那么这些词是动态概念，不作 l0 概念。

例如：第一图像(110)给//l0 消费者一种安全感。

句中用“//”加具体语块或概念的形式，标明其与规则的对应。

- 位置信息

逻辑概念都可以用于三主块句，当三主块句的 EG 位于句尾时，我们优先选择这类词为 l0 概念，规则如下：

(0)CHN[把,将,对,向,由,比,与]+(f)LC_CHK[EG]&LC_E_SCORE[E_TAIL]&LC_SC_KEY[3KUAI]
=>LC_SELECT(0,LC_CC,l0)$

例如：移动终端对//l0 信号能量进行探测//EG。

- 个性特征

对于每个 l0 概念的个性特征，将分别制定排除规则。以“对”为例，《现代汉语词典》(第六版)中，“对”共有 16 个义项，对应于 HNC 概念有 5 个概念，分别是动态概念（v）、值概念（zzp，zzw）、主语块标识符（l0）、静态概念（g）、伦理属性概念（jgu841），如下表所示：

义项	词性	HNC 解释	概率类别	对应《现代汉语词典》	示例
1	动词	动态概念	v	义项 1、2、3、4、5、6、7、8、9、11、12	～答、～立、～质、～半儿
2	量词	值概念	zzp，zzw	义项 14	一～鹦鹉
3	介词	主语块标识符	l0	义项 15	～他表示谢意
4	名词	静态概念	g	义项 13	喜～
5	形容词	伦理属性概念	jgu841	义项 10	你的话很～

表 4：词语“对”的概念特征

“对”需要排除的是作量词（值概念 zzp,zzw）、作形容词（伦理属性概念 jgu841）和作介词（辅语块标识符 l1）的情况。可通过以下规则排除：

(-1)LC_CC[j30,l91]+(0)CHN[对]=>!LC_SELECT(0,LC_CC,l0)$ 当“对”前面有数词或指示代词

时，“对”为量词。

(0)CHN[对]&END%=>!LC_SELECT(0,LC_CC,l0)$ 当“对”位于句尾时，“对”不作 l0 概念。

(0)CHN[对]+(1)CHN[的,了]&END%=>!LC_SELECT(0,LC_CC,l0)$ 当“对”后面是“的”或“了”且位于句尾时，“对”不作 l0 概念。

(0)CHN[对]+(f){CHN[来说,说来,而言]}=>!LC_SELECT(0,LC_CC,l0)$ 当“对”后面有“来说，说来，而言”时，“对”为辅块标识符 l14。

5．2 识别规则

识别规则主要用来识别格式转换是发生在主句还是小句（从句）中。这是进行下一步句子分析和语序调整的依据。我们在逻辑概念 10 上标记 level 属性，用以表明逻辑概念的级别。level=1 表示 10 是全局的语块标识符，level=2 表示 10 是局部的语块标识符，数字越大，表示级别越低。

识别规则阶段，除了切分了小句，识别出 EG、ABK、LB、10 概念之外，没有其他的信息可供利用。所以要识别出 LEVEL=1 的 10，需利用知识库中 10 的句类信息和 EG 的句类信息。如果 EG 前面的 10 的句类信息和 EG 的句类信息匹配，那么这个 10 的 LEVEL 等于 1：

(0)LC_CC[l0]&LC_SC[XY]+(f){LC_CHK[EG]&LC_SC[XY]}=>PUT(0,LEVEL,1)$

例如：播放器对//10 该内容进行解扰(208)//EG。

10“对”的句类可以是作用句（X），EG“进行解扰”的句类也可以是作用句（X），它们的句类信息相匹配，所以此处 10 的 LEVEL 是 1。

5．3 转换规则

采用排除规则可以排除不进行格式转换的句子，采用分析规则可以识别出 l0 的层次。识别过程结束后，会产生一颗分析树，转换规则将在这棵树上进行，如下图所示：

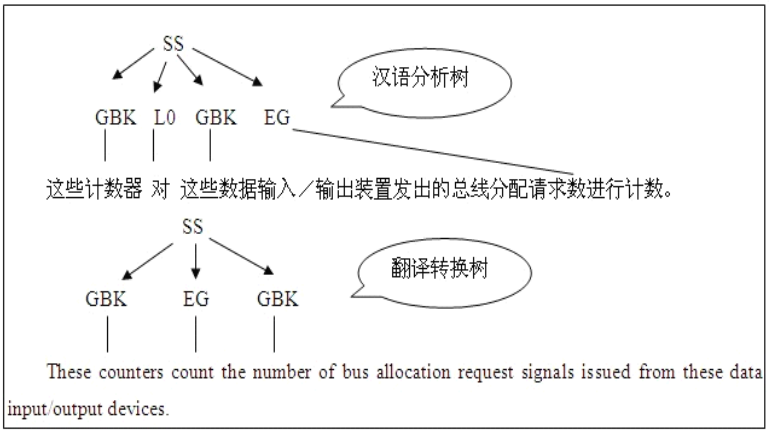

图 1：格式转换分析树

以由“对”所构成的格式为例，格式转换规则总的来说有以下特征：

“对”，可用于反应句、信息转移句、交换句、一般承受句、因果句、一般判断句、约束句、单向关系句、作用句、关系自身转移句、效应句，主要作为 GBK2 的标识符 l02。可用于三主块句也可用于四主块句，通常采用!11、!113 格式。不管用于四主块句还是三主块句，其 EG 都不带下装（hv）。

当“对”用于三主块句时，EG 通常为高低搭配 EQ+E 结构，如“进行描述、进行解扰、进行计数”等。

- 当句子中存在 GBK1（即主语不缺省）时，翻译成英语时采用主动格式。

规则如下：

(-1)LC_CHK[GBK]+(0)CHN[对]&LC_CC[l0]&LC_LEVEL[1]+(1)LC_CHK[GBK]+(2)LC_CHK[EG]=>(-1)+DEL_NODE(0)+(2)+(1)$

例如：这些计数器//GBK 对//l0 这些数据输入 / 输出装置发出的总线分配请求数//GBK 进行计数//EG。（ These counters count the number of bus allocation request signals issued from these data input/output devices.）

- 当句子中没有 GBK1（即主语缺省）时，翻译成英语时采用被动格式。

规则如下：

(b){!(LC_CHK[GBK])}+(0)CHN[对]&LC_CC[l0]&LC_LEVEL[1]+(1)LC_CHK[GBK]+(2)LC_CHK[EG]=>DEL_NODE(0)+(1)+(2){VOI=P}$

例如：以上结合本发明的优选实施方式对//l0 本发明//GBK 进行了描述//EG。（ The present invention has been described above in connection with the embodiments of the invention.）

当“对”用于四主块句时，翻译成英语需要在 GBK2 前面加介词（如 to、for 等）。

- 当句子中存在 GBK1（即主语不缺省）时，翻译成英语时采用主动格式，并在位置(1)前增加介词 to/for。

规则如下：

(-1)LC_CHK[GBK]+(0)CHN[对]&LC_CC[l0]&LC_LEVEL[1]+(1)LC_CHK[GBK]+(2)LC_CHK[EG]+(3)LC_CHK[GBK]=>(-1)+DEL_NODE(0)+(2)+(3)+ADD_NODE{EGN=[for,to]}+(1)$

例如：第二通信模块//GBK 对//l0 计算机系统//GBK 提供//EG 第二格式的第二表示数据//GBK。（ The second communications module transmits the second indicating data in a second format to the computer system.）

- 当句子中没有 GBK1（即主语缺省）时，翻译成英语时采用被动格式。

规则如下：

(b){!(LC_CHK[GBK])}+(0)CHN[对]&LC_CC[l0]&LC_LEVEL[1]+(1)LC_CHK[GBK]+(2)LC_CHK[EG]+(2)LC_CHK[GBK]=>DEL_NODE(0)+(3)+(2){VOI=P}+(1)$：

没有 GBK1 时，翻译成英语采用被动格式。

例如：在持久操作期间，尽管电池包耗尽，仍可对//l0 便携式终端//GBK 稳定地提供//EG 电源//GBK。（ The power can be stably provided to the portable terminal in spite of depletion of a battery during a long-duration operation.）

6 实验结果分析

我们随机抽取了 3000 个句子对排除规则和 LEVEL=1 的转换规则进行了人工评测，评测结果能达到 85%的准确率。

对评测结果进行分析，发现问题主要集中在以下几方面：

分词的影响。如“则由轨迹结构对调焦误差信号的调制最小。”句中，“对调”被切成了一个词。

EG 规则的影响。如“将由数据排序装置所排序的数据中的有效数据输出到装置外部,”中，当“将”后面有“由”时，EG 识别制定的规则是“将”为 QE。

辅块规则的影响。如“反射区域内液晶分子与聚合物的比比透射区域内低。”中，第二个“比”被当成了 l1。

EG 识别的影响。如“本发明所述方法对 MPLS LSP 的性能参数测量做了详细的规定。”中，将“规定”识别为了 E，因而影响了 l0 概念“对”的识别。

7 总结与展望

本文针对汉英专利格式转换语料标注了转换现象、依据、结果和规则。定义了汉英专利格式转换的规则符号，对规则进行了形式化表示。总结了汉英专利格式转换的规则，包括排

除规则、识别规则和转换规则。并对转换规则进行了人工评测，取得了较好的实验效果。

下一步的工作是，继续对排除规则、识别规则和转换规则都进行人工评测，针对性改进规则，提高规则效果；同时改进程序，提高程序的处理效果；另外，还需进一步扩大研究范围和研究深度，将格式转换的研究范围扩大到所有语言逻辑 l0 概念；并且研究格式转换发生在语块内部的情形。

参考文献：

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

[2] 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.

[3] 张艳红.英汉互译中的格式转换[C]//张全，萧国政.HNC 与语言研究.湖北：武汉理工大学出版社，2001，302-307.

[4] Joseph H. Greenberg，William Croft. Genetic Linguistics: Essays on Theory and Method [M]. USA：Oxford University Press，2005.

[5] 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.

[6] 孙广范.句法调序的统计机器翻译方法研究[J].计算机工程与应用，2009，45（36）：142-144.

[7] 孙雄勇.汉英翻译中一般转移句格式转换[C]//苗传江，杜燕玲.第二届 HNC 与语言学研讨会论文集.北京：海洋出版社，2004，362-367.

[8] 曾维，张克亮.汉英翻译中一般转移句格式转换[C]//苗传江，杜燕玲.第二届 HNC 与语言学研讨会论文集.北京：海洋出版社，2004，362-367.

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

[10] 连巍巍，张克亮.面向汉英机器翻译的格式自转换研究[C]//朱小健，张全，陈小盟.HNC 与语言学研究（第 4 辑）.北京：北京师范大学出版社，2010，297-303.

[11] 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.

作者联系方式：刘智颖　北京师范大学中文信息处理研究所 100875　13521008057 liuzhy@bnu.edu.cn