

# 面向汉英机器翻译的专利文献小句变换研究

王立霞

(北京师范大学 中文信息处理研究所, 北京 100875)

**摘要:**专利文献的自动翻译是机器翻译的一个重要应用领域,复杂长句的翻译是汉英机器翻译的难点。本研究期望找出汉英复杂长句中从句变换的形式化转换规则。汉语复杂长句中会包含多个小句,这些小句都是独立存在的,但翻译成英语时,一般只有一个核心小句,其他小句都变换成 doing、to do、从句或短语等其它形式。文中以 1300 句汉英双语专利文献语料为研究对象,对汉语中的小句翻译为英语的变换情况进行分类研究,从小句句间关系、共享关系的角度出发,描述激活特征,并按五种变换方式分类,提出了十二条变换规则,小规模语料实验结果证明规则可行有效。下一步工作需要扩充研究语料,对语料进行更深入的挖掘和分析,在更大规模语料中验证规则的实用性。

**关键词:**机器翻译;小句;小句变换

**中图分类号:**TP31

**文献标识码:**A

**文章编号:**1673-629X(2012)11-0077-04

## A Chinese-English MT-Oriented Study on Small Sentence Pattern Transformation of Long Patent Sentence

WANG Li-xia

(Institute of Chinese Information Processing, Beijing Normal University, Beijing 100875, China)

**Abstract:** Automatic translation of the patent document is an important field of machine translation, and the translation of long patent sentence is the difficulty in Chinese-English machine translation research. A complex sentence contains more than one small sentence, which is independent in Chinese. There is only one core sentence when translated into English from Chinese. Other sentences are transformed into clauses or non-defining phrases. It is expected to find the transformation rules that how Chinese small sentence translated into English. Study 1300 Chinese-English bilingual sentences, from the standpoint of small sentence pattern transformation, according to semantic relationship and sharing relationship among small sentences, describe the activation characteristics, sum up 12 formal sentence analysis rules and transformation rules. The small scale experiment results indicate these rules are effective. The next step needs to expand the research corpus and verify the practicality of these rules in larger corpus.

**Key words:** machine translation; small sentence; small sentence pattern transformation

### 1 问题提出

汉英机器翻译系统应跨越逗号,以句号为翻译单位。但汉语逗号的使用过于宽泛,一个句号之内包含多个逗号,甚至一逗到底,一个句号成为一段,这样的复杂长句成为汉英机器翻译的一大难点。

专利文献的自动翻译是机器翻译的一个重要应用领域。专利文献的特点是一项权利要求要用一句话表述出来,因此复杂长句在专利文献中尤为明显。专利文献机器翻译必须要正确处理复杂长句这一难题。

HNC 理论<sup>[1,2]</sup>将以句号或与句号等价的问号、感叹号等为结束标记的文本片段称为语段,在机器翻译研究中又称为句群。句群内由逗号分割的部分称为语

串,约 65% 的语串成句,如果这些语句是句群的构件,则称为小句,其余可统称为非小句语串。

汉语复杂长句中会包含多个小句,这些小句都是独立存在的,但翻译成英语时,一般只有一个核心小句,其他小句都变换成 doing、to do、done、从句或短语等其它形式。这种变换称为小句变换。

以下示例是从专利文献语料中抽取的以句号为单位的句子,它们翻译成英语都存在小句变换的情况。句 A 为人工给出的参考译文,句 B 是 google 翻译系统中给出的结果。

例 1:通信装置(27)最好<是:eg>一个移动电话,<+>它能够<打电话:eg-which>到通信网络。

A: The communication device (27) is preferably a mobile telephone which can telephone a communications network.

B: The best communication device is a mobile

phone, it can call the communications network.

例 2: 本发明<涉及: eg>一种具有一定刚度的和脆性的闪光金属片, <+>它的中间层<是: eg-sl>反射材料, ++由位于两个表面的绝缘层<支撑: eg-done>。

A: A rigid and brittle bright metal flake is formed of a central layer of a reflective material supported on both sides by dielectric layers.

B: The present invention relates to a certain stiffness and brittleness of sheet metal flashing, it is the middle layer of reflective material, from the surface of the insulating layer in the support of the two.

例 3: 本发明片剂的特征<在于: eg>此润滑剂为粉末的形式, 至少大部分<分布: eg>在此片剂的表面, <而且: lb1/2-and><根据: l11>《法国药典》(第 10 版, V. 5. 1-“片剂的易碎性”, 1993 年 1 月)中说明的方法进行<测定: eg-done>+, 其易碎性<小于: eg>1%。

A: The invention is characterized in that the lubricating agent is in powder form and is distributed at least for the greater part at the tablet surface and its friability measured as specified in the French codex (10th edition, V. 5. 1- Friability of tablets, January 1993) is less than 1%.

B: The characteristics of tablets of the invention this lubricant in the form of powder, at least most of the distribution in the tablet surface, and according to "the French Pharmacopoeia" (10th edition, V. 5. 1-"the fragility of tablets," January 1993) the method described in the determination of less than 1% of its fragility.

从上述例子可见, 现有汉英机器翻译系统对小句句间关系考虑不足, 汉语翻译为英语时没有进行应有的小句变换处理。

## 2 前人成果

已有学者<sup>[3]</sup>对机器翻译研究的新进展进行了总结, 有学者<sup>[4]</sup>从句类及句式转换的角度对汉英机器翻译进行了研究, 有些学者<sup>[5-7]</sup>对专利文献汉英机器翻译进行了专门研究, 统计分析显示共有 39% 的句子包含多个小句, 其中有 85% 的句子需要进行小句变换处理。

宋柔<sup>[8]</sup>对现代汉语跨标点的句法关系进行了研究, 提出了现代汉语中标点句和跨标点句的句法关系的概念, 设计了表示跨标点句句法关系的换行缩进的直观表示方法, 揭示了跨标点句句法关系的性质, 包括句法关系类型的搭配条件、栈式结构规律以及栈式结构进退的限度。

侯敏、孙建军<sup>[9]</sup>从零形回指的角度研究了小句句

间关系, 详细分析了汉语零形回指的确认、类型、产生的原因及使用的条件, 指出其对汉英机器翻译造成的主要障碍是生成的英语句子在结构上不合语法, 并提出在句组层面上解决问题的算法。

张全、吴晨、韦向峰<sup>[10]</sup>分析了语句之间语义块共享的类型, 给出了语句间语义块共享的具体分类, 统计了真实语料中各共享类型的分布数据。

贾宁、张全<sup>[11]</sup>在句类分析的基础上, 从小句间语义块共享关系的角度分析语义块的省略。将语义块的省略分为语义块整块共享形成的省略和语义块部分共享形成的省略, 并给出了相应的处理算法。

池毓焕、李颖<sup>[12]</sup>提出大句范式的概念, 在大句的范围内讨论了某些小句变换的模式。

黄河燕、陈肇雄<sup>[13]</sup>提出了基于多策略分析的复杂长句翻译处理算法, 综合利用源语言句子中多种相关的语言特征, 包括语法语义特征、句子长度、标点符号、功能词以及上下文语境条件等, 对复杂长句进行切分简化处理和译文的复合生成。

这些研究虽然涉及到汉英机器翻译中的小句句间关系识别及小句变换, 但研究还不深入、全面, 特别是对小句变换处理, 还没有提出明确的处理办法。

## 3 小句变换激活特征

汉语长句中小句都是独立存在的, 小句之间具有并列、转折、递进等语义关系, 同时由于语言交际中的经济原则, 小句之间还可能存在着语言成分共用的现象, 称为共享关系。

文中对 1300 句汉英双语专利文献语料进行了分析, 从中发现有 398 句存在小句变换情况, 约为语料的 30.62%。逐一分析了这些小句, 从小句间语义关系及有无共享关系角度出发, 总结出如下变换激活特征。

### 3.1 小句间无共享关系

1. 有明显的关联词, 即存在复句关系, 如: 并列、转折、递进等示例: 首先, 口袋隔离膜必须能从其本身剥离开来, 而且它又必须不粘附到背衬膜的压敏粘合剂上。

参考译文: First, the pouch barrier film must be peelable from itself, yet it must not adhere to the pressure sensitive adhesive on the backing film.

2. 无关联词, 小句间的 GBK1、GBK2 存在关联 (GBK1 相当于传统语法中所说的主语, GBK2 相当于传统语法中所说的宾语, 下同)。

(1) 第二个小句的 GBK1 重复第一个小句的 GBK2。

示例: 形成一个在导电构件与地之间连接的电流路径 A, 该电流路径 A 不经过电子源和驱动电路中的

任何一个。

参考译文:An electric current flow path A is formed as extending between the electroconductive member and the ground without passing through any of the electron source and the drive circuit.

(2)第二个小句的 GBK1 为代词,指代第一个小句的 GBK2。

示例:通信装置(27)最好是一个移动电话,它能够打电话到通信网络。

参考译文:The communication device (27) is preferably a mobile telephone which can telephone a communications network.

(3)第二个小句的 GBK1 为第一个小句的 GBK2 的一部分。

示例:一个综合控制器,它包括多个控制器,诸控制器中的每一个都执行一个不同的程序。

参考译文:An integrated controller includes a plurality of controllers, each of the controllers executing a different program.

3. 无关联词,小句间的语义关系明确。

(1)第二小句为第一小句的目的。

示例:该散列密钥与本地存储的散列密钥列表相比较,看看先前是否已经备份了该本地文件。

参考译文:The hashing key is compared to a list of hashing keys stored locally to see if the local file has been previously backed up.

(2)第二小句为第一小句的结果。

示例:食用磷酸能和烟草中生物碱结合形成黄色,在烟草薄片生产过程中不需要加入黄色的色素。

### 3.2 小句间有共享关系

1. 共享主块(GBK1/GBK2/GBK3),有关联词。

A 并列关系。

示例:上述接触板(5;12;15;24;29;35)或者牢固固定在安装支座(2;31)上,或者可相对于外壳进行有限轴向运动。

参考译文:The contact plates (5; 12; 15; 24; 29; 35) can either be securely fixed in the mounting foot (2; 31) or have limited axial movability in relation to the housing.

B 递进关系。

示例:尤其是,L 蛋白不仅可分泌到细胞外,而且可在细胞内积聚。

参考译文:In particular, the L protein is not only excreted outside the cell but also accumulated within the cell.

2. 共享主块(GBK1/GBK2/GBK3),无关联词。

示例:该制品包括一个单层长形导体,具有相反的两端。

参考译文:The article comprises a single layer elongate conductor having opposing ends.

示例:当第二通信终端空闲时,向第一通信终端发送一个消息,指示现正处于空闲状态。

参考译文:When the second telecommunication terminal becomes idle, the second terminal sends a message to the first terminal indicating the idle state.

## 4 小句变换类型及规则

通过对上述小句变换语料的分析,将小句变换分为以下五个类型,并初步总结出一些变换规则:

(一)小句的核心动词变为 ing 形式:eg-doing。

规则一:第二小句共享第一小句的 GBK2,则第二小句的 eg 可变为 ing 形式。

示例:当第二通信终端空闲时,向第一通信终端发送:eg>一个消息,+<指示:eg-doing>现正处于空闲状态。

参考译文:When the second telecommunication terminal becomes idle, the second terminal sends a message to the first terminal indicating the idle state.

规则二:第二小句的 GBK1 重复第一小句的 GBK2,则第二小句的 eg 可变为 ing 形式。

示例:此外,RAID 设备还向每一个队列对(51-54)<发送:eg>一个描述符,<+>每个描述符<表示:eg-doing>数据块的 N 分之一。

参考译文:Further, the RAID device posts a descriptor to each of the queue pairs, with each descriptor referring to 1/Nth of the data block.

规则三:后一小句首以表示因果、目的关系的连词(非括号型连词)开头,则该小句 eg 通常变为 ing 形式。

示例:这样,<既:lb1/2><改善:eg>扫描电极的选择比,<又:lb1/2>最大限度地<减少:eg>波形差异的产生,<从而:lb2/2-thereby><减少:图像中产生的串扰。

参考译文:Therefore, the selection ratio of scanning electrodes are improved and the generation of the waveform differential is minimized, thereby reducing the crosstalk generated in a picture.

规则四:第二小句首为“同时”+eg 形式,则该小句的 eg 通常变为 ing 形式。

示例:因此,生产率<提高:eg>了,同时<节省:eg-doing>了劳力<并:l>更准确地<检查:eg>出次品。

参考译文:Therefore, productivity is increased while saving labor and inferior products can be accurately detected.

ted.

(二)小句的核心动词变为 to do 形式:eg-to do.

规则一:第二小句以表目的的“使”字开头,则该小句的 eg 通常可变为 to do 形式。

示例:平衡杆的设计<降低:eg>了其第二弯曲方式响应,++<使:eg-to do>其具有一个可以比流动管共振驱动频率低的频率。

参考译文:The balance bar's design lowers its second bending mode response to have a frequency that may be lower than the flow tube resonant drive frequency.

规则二:小句中有“用于”、“用以”、“以”等表示目的的 11 概念,则该小句的 eg 可变为 to do 形式。

示例:调制器(106)的另一个区域<有:eg>几条线,++<用于:117>对印刷的图象<实现:eg-to do>行组合,<以:1b2/2><产生:eg-doing>更多的灰度级。

参考译文:Another area of the modulator (106) has lines designated to perform row integration on the print image, allowing for more gray levels.

规则三:第一小句的 eg 在句尾,第二小句从语义上看为第一小句的目的,则第二小句的 eg 可变换为 to do 形式。

示例:该散列密钥与本地存储的散列密钥列表相<比较:eg>,<看看:eg-to do>先前是否已经备份了该本地文件。

参考译文:The hashing key is compared to a list of hashing keys stored locally to see if the local file has been previously backed up.

(三)小句的核心动词变为从句形式:eg-which eg-that.

规则一:后一小句以“使”字开头,则该小句 eg 可以变换为从句形式。

示例:在又一个实施例中,对着每只手的静电发射器<产生:eg>脉冲,<使:eg-that>每个脉冲具有相反的电荷。

参考译文:In still another embodiment, the electrostatic emitter directed at each hand is pulsed such that each pulse has an opposite charge.

规则二:第二小句 GBK1 以代词形式重复第一小句 GBK2,则第二小句可变换为 which 从句。

示例:通信装置(27)最好<是:eg>一个移动电话,<+>它能够<打电话:eg-which>到通信网络。

参考译文:The communication device (27) is preferably a mobile telephone which can telephone a communications network.

规则三:第二小句的 GBK1 共享第一小句的 GBK2,则第二小句可变换为 which 从句。

示例:本发明<公开:eg>了一种用于燃料电池系统的重整器,++<包括:eg-which>具有催化剂层空腔的反应器主体。

参考译文:The invention discloses a reforming device of fuel battery system, which contains reactor bulk with cavity of catalyst layer,

(四)小句的核心动词变为 done 形式:eg-done.

规则一:当后一小句为规范格式中的! 12 格式时,后一小句的核心动词可变换为 done 形式。

示例:本发明<涉及:eg>一种具有一定刚度的和脆性的闪光金属片,<+>它的中间层<是:eg-sl>反射材料,++由位于两个表面的绝缘层<支撑:eg-done>。

参考译文:A rigid and brittle bright metal flake is formed of a central layer of a reflective material supported on both sides by dielectric layers.

(五)小句的核心动词省略:eg-sl.

规则一:当两个小句存在相同的核心动词,后一小句补充说明第一小句内容时,后一小句的核心动词可以省略。

示例:本发明<涉及:eg>一种农业上应用的化学农药,<特别是:cry><涉及:eg-sl>一种防治烟草蚜虫的制剂。

参考译文:The invention relates to a chemical pesticide used in agriculture, in particular to a preparation for preventing and treating the tobacco aphid.

## 5 结束语

文中对专利文献汉英机器翻译中的小句变换进行了初步研究,小句变换所需的激活特征还很粗浅,变换规则需要进一步细化和形式化,对没有明显语言标记的小句间关系识别需要加强。下一步工作需要扩充研究语料,对语料进行更深入的挖掘和分析,在更大规模语料中验证规则的实用性。

### 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] 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.
- [3] 刘群. 机器翻译研究新进展[J]. 当代语言学,2009,11(2):147-158.
- [4] 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.
- [5] Jin Yaohong. Improving Chinese-English patent machine translation using sentence segmentation[C]//Proceedings of the 6th International Conference on Natural Language Pro-

(下转第84页)

果图,图中显示的均为经过畸变校正与平行校正后的输出图像。左半部分为左摄像机拍得的图像,右半部分为右摄像机拍摄的图像。

由以上 4 种匹配方法得出的性能结果统计见表 1。由匹配结果可以得出,采用小波变换减少了匹配点数,降低了匹配复杂度,减少了算法运行时间;而采用极线约束的方法,在同一水平线上搜索匹配点,减少了误匹配点数,提高了正确匹配率。综合采用小波域 SIFT 变换和极线约束相结合的算法在降低了特征提取和特征匹配的复杂度的同时,提高了图像中特征点的正确匹配率,同时大大增强了 SIFT 算法的实时性。

表 1 匹配结果统计

算法	SIFT	小波+ SIFT	SIFT+极 线约束	文中方法
总时间(s)	24.234	6.047	25.5	5.41
匹配点数	435	81	401	79
误匹配点数	36	5	23	3
匹配率	91.72%	93.82%	94.26%	96.20%

### 3 结束语

从上述理论分析和实验结果可以看出,基于小波域和 SIFT 算法的方法能够满足较高精度,同时使得特征提取和匹配的复杂度大大降低,SIFT 算法的实时性和准确性得到提高。

由以上算法分析和实验结果可见:

(1)文中提出的算法采用基于 SIFT 的特征匹配,在保证匹配结果有效性和准确性的同时,极大提高了匹配结果对图像噪声和图像变换的鲁棒性和抗干扰性。

(2)SIFT 算法检测特征点时采用高斯核和 DOG 算子对图像进行多次运算,耗费了大量时间。文中算法在生成 SIFT 特征描述之前先对图片进行小波分解,可减少特征描述生成的次数,使得特征描述生成阶段的计算量减少,大大提高了实时性。

(3)本算法采用了极线约束,可以降低 SIFT 算法的搜索范围,从而可以在算法运行时间减少的同时使得算法准确性得到提高。

(4)文中算法生成的特征点个数远远少于 SIFT 算法,从而待匹配特征点数和数据库容量得到减少,因而匹配时间也缩短了。

### 参考文献:

- [1] 游素亚,徐光祜. 立体视觉研究的现状与进展[J]. 中国图象图形学报,1997,32(2):17-23.
- [2] Yoon Kuk-Jin, Kweon In-So. Locally Adaptive Support-weight Approach for Visual Correspondence Search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006,28(4):650-656.
- [3] Lowe D G. Object recognition from local scale-invariant features[C]//International Conference on Computer Vision. Corfu, Greece:[s. n.],1999.
- [4] 全 斌. 数字图像特征点提取及匹配的研究[D]. 西安:西安科技大学,2009.
- [5] Mallat S G. A theory of multi resolution signal decomposition: the wavelet representation signal decomposition[J]. IEEE-PAMI,1989,11(7):647-693.
- [6] 刘佳嘉,何小海,陈为龙. 一种结合小波变换的 SIFT 特征图像匹配算法[J]. 计算机仿真,2011,28(1):257-260.
- [7] 赵钦君,赵东标,韦 虎. Harris-SIFT 算法及其在双目立体视觉中的应用[J]. 电子科技大学学报,2010,39(4):546-550.
- [8] 孙延奎. 小波分析及其应用[M]. 北京:机械工业出版社,2005.
- [9] 谢 凡,秦世引. 基于 SIFT 的单目移动机器人宽基线立体匹配[J]. 仪器仪表学报,2008,29(11):2247-2252.
- [10] 宰小涛. 基于 SIFT 特征描述子的立体匹配算法研究[D]. 上海:上海交通大学,2007.
- [11] 王 艳. 基于点特征的立体匹配算法研究[D]. 南京:南京理工大学,2009.
- [12] 王永明,王贵锦. 图像局部不变性特征与描述[M]. 北京:国防工业出版社,2010.

(上接第 80 页)

- cessing and Knowledge Engineering. [s. l.]:[s. n.],2010.
- [6] Wang Dan. Chinese to English automatic patent machine translation at SIPO[J]. World Patent Information,2009,31(2):137-139.
- [7] Goto I, Lu Bin, Chow K P. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop[C]//Proceedings of NTCIR-9 Workshop Meeting. Tokyo, Japan:[s. n.], 2011:559-578.
- [8] 宋 柔. 现代汉语跨标点句句法关系的性质研究[J]. 世界汉语教学,2008(2):26-44.
- [9] 侯 敏,孙建军. 汉语中的零形回指及其在汉英机器翻译中的处理对策[J]. 中文信息学报,2004,19(1):14-20.
- [10] 张 全,吴 晨,韦向峰. 汉语句间成分共享类型及分布研究[J]. 计算机科学,2007,34(1):166-169.
- [11] 贾 宁,张 全. 基于句间关系的汉语语义块省略恢复[J]. 中文信息学报,2008,22(6):33-37.
- [12] 池毓焕,李 颖. 面向汉英机器翻译的大句范式初探[C]//中国计算机语言学研究前沿进展(2007-2009). 烟台:出版者不详,2009:395-400.
- [13] 黄河燕,陈肇雄. 基于多策略分析的复杂长句翻译处理算法[J]. 中文信息学报,2002,16(3):1-6.