

机械制造工艺汉英自动翻译系统的研究与实践

周继如

(南京师范大学电气与电子工程学院, 南京, 210042)

[摘要] 从机械制造工艺语言的特点入手, 对机械制造工艺汉英自动翻译系统的相关问题进行了研究, 在理论研究的基础上, 结合成都飞机公司的具体要求和情况, 开发研制了飞机工艺文件编制及汉英自动翻译系统。

[关键词] 制造业全球化; 工艺语言; 自动翻译

[中图分类号] TH16; H085; [文献标识码] A; [文章编号] 1008-1925(2001)01-0044-07

0 引言

从第一次工业革命兴起到目前的200多年间, 机械制造技术发生了很大的变化。由于机械科学、计算机科学、电子科学、系统科学和制造科学的发展, 机械制造在近几十年中已形成向精密化、自动化、智能化、分散化与网络化、敏捷化、全球化发展的趋势。随着制造工业的全球化, 以及敏捷制造、并行工程的发展, 不同国度的企业之间的生产合作日益增加, 产品制造过程中的交流也越来越多, 异地设计、异地制造的现象十分普遍。其中产品设计采用的是全世界通用的图形语言, 而工艺文件采用的却是各不相同的文字语言, 这就给工艺上的沟通带来很大的不便, 工艺文件的翻译成为一个亟待解决的问题。

由于机械制造中零件种类多、工艺复杂, 造成工艺文件中词汇量大且专业性强, 语句错综复杂、样式繁多, 传统的人工翻译必须投入大量的人力、物力和财力。不仅投入大, 而且周期长, 翻译结果难以保证一致性, 远远满足不了规模化生产的要求, 这就必须在翻译技术上革新。制造工艺语言的自动翻译(automatic translation)应运而生。

在我国, 随着改革开放的深入以及“入世”在即, 国内机械制造企业与国外的合作尤为频繁, 异地设计、异地制造的现象十分普遍。所以制造工艺文件的汉外——外汉自动翻译更是一个迫切需要解决的问题。而制造工艺语言的自动翻译在国内自动翻译领域中尚属空白。

本文正是面向企业需求, 对机械制造中工艺文件的汉英自动翻译的相关问题进行研究, 并在理论研究的基础上, 结合成都飞机公司的超七飞机的具体要求和情况, 开发研制了飞机工艺文件编制及汉英自动翻译系统。

1 机械制造工艺语言的特点

机械制造工艺语言在词汇和句型上有自己的特点, 作者在词汇和句型两方面对机械制造工艺语言进行了深入地研究。

* 收稿日期: 2000-09-05

作者简介: 周继如, 1971-, 女, 南京航空航天大学机电控制及自动化硕士研究生, 现为南京师范大学电气与电子工程系讲师, 主要从事机电一体化方面的研究。

1.1 词汇

在词汇上,为了更好地实现翻译自动化,词的合理分类是关键.根据汉语语言学的理论,并充分考虑到工艺语言的特点,作者将词分为名词、动词、形容词、数量词、副词、介词、搭配词、时态助词、结构助词、连词十大类.每种类型的词还可以取子值,如:

- (1) 名词次类: 一般名词、专用名词、人称代词、物主代词、关系代词.
- (2) 动词次类: 情态动词、联系动词、助动词、一般动词.
- (3) 词组类型: 名词词组、形容词词组、前置词词组、后置词词组、动词句、从句.....

但是由于汉语没有形态变化的特点,即使将词合理分类后,在工艺文件中,还是会出现同形歧义现象,如下表所示.

表1 工艺文件中的同形歧义现象

序号	同形词	例	句
1	名词与动词同形	零件的加工	加工外圆
2	形容词与方位副词同形	内螺纹	在夹具内
3	介词和连词同形	和模具样板	外圆和端面

为了解决这些同形歧义现象,需要制定相应的消除兼类的规则库.

1.2 句型

在句型上,工艺文件中采用的语句基本为陈述句.通过对机械制造工艺文件中汉语语句的分析,作者大概归纳出如下表所示的几种基本句型:

表2 工艺文件中的基本句型

序号	基本句型	例	句
1	(状语+) 谓语	按简图加工.	
2	谓语+ 宾语	磨削零件.	
3	(状语+) 谓语+ 宾语	按兰图车削外圆和端面.	
4	谓语+ (定语+) 宾语	检查零件的合格检印.	
5	(定语+) 主语+ 谓语+ 宾语	对合处的间隙不小于0.5.	
6	主语+ 谓语+ (定语+) 宾语	接头是特别重要的零件.	
7	"把" "将" 字句	把缝焊好的油箱置于入置架上.	
8	"被" 字句	出油口被油塞堵上.	
9	省略介词句	(用) 虎钳夹紧零件.	
10	被动句	在端头四个压紧器要压紧.	
11	兼语式	保证尺寸不超过要求.	
12	复合句	先将链带与链带环节装好,而后用链带环节组成链条,再装链条并收紧.	

2 机械制造工艺语言的描述和运算模型

2.1 工艺语言描述的数学模型

由于在自动翻译中都是以每个单词的词类而不是具体的单词作为语法分析对象的,所以工艺语言自动翻译系统中描述工艺语言的数学模型主要指词类的模型,包括两个方面:

2. 1. 1 代码体系

1——名词;2——动词;3——时态助词;4——数量词;5——副词;6——形容词;7——介词;8——搭配词;9——结构助词;10——连词

2. 1. 2 集合的定义

前导集——对一给定的词性其左邻的所有可能的词性的代码的集合.

后导集——对一给定的词性其右邻的所有可能的词性的代码的集合.

2. 2 工艺语言运算的数学模型

系统采用属性制约文法进行运算,为此建立了属性制约文法的数学模型.如属性制约文法对词类的判别方法是按下述原则进行的:句子中的各个词在该句中的词性是合理的充分必要条件是该句中任意一组相邻的三个词应满足,中间词的词性应在前一词的后导集与后一词的前导集的交集之中.

所以建立每次取 3 个词的数学模型如下: 设 m 为句长, 对 $I = 1, 2, \dots, m - 2$ 计算

$$J[I, I + 1] = Cq[I] \cap Cp[I + 1]$$

$$J[I + 1, I + 2] = Cq[I + 1] \cap Cp[I + 2]$$

$$Cq[I + 1] = \{Unq\}(c[i] \cap J[I, I + 1])$$

$$Cp[I + 1] = \{Unp\}(c[j] \cap J[I + 1, I + 2])$$

$$Cq[I + 2] = \{Unq\}(c[j] \cap J[I + 1, I + 2])$$

若对每一个 $I = 1, 2, \dots, m - 2$ 有 $J[I, I + 1] \cap J[I + 1, I + 2] \neq \emptyset$ 则有解, 否则无解(可能为非线性结构).

另外, 作者对生成汉语语法树以及语法树的转换等都建立相应的运算模型.

3 系统的总体框架

工艺语言汉英自动翻译系统的基本策略是基于树形图的转换法. 基本原则是结合工艺语言的特点, 尽量发挥语法分析的作用辅以语义搭配判断. 由于工艺文件中基本句型是陈述句, 所以语法基础是可以诠释所有汉语陈述句的汉语完全语法树和属性制约文法. 本系统由原语(汉语)的线性结构出发, 经过多层次、多次数的扫描, 按规则的有序匹配, 形成以动词为根结点, 以逻辑语义项为主结点的多结点、多标记的树形结构, 最后从根结点逐层展开, 形成目标语言(英语)的线性结构, 得到相应的译文. 系统的总体结构如图 1 所示.

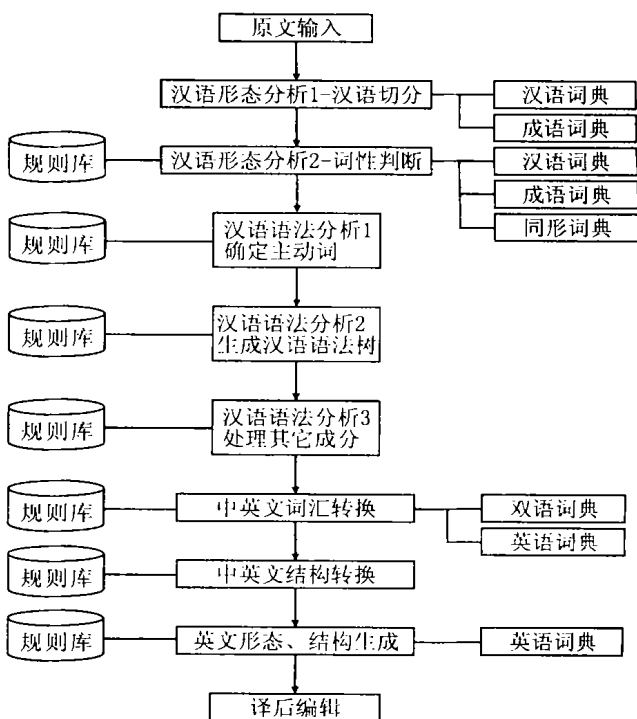


图 1 系统总体结构图

4 工艺机器词典的编制

在工艺语言汉英自动翻译系统中, 根据语法、语义分析的需要, 建立了6部机器词典, 包括: 1 部汉语词典、1 部成语词典、2 部语法多义词典、1 部双语词典和 1 部英文词典。

4.1 汉语词典

汉语词典不仅储存了单义词的基本词义和词类信息, 而且还带有它们在以后程序中需要进一步确定词义和语法功能的加工信息。比如: 该词是否可以构成成语、是否能组成独特结构、是否为多义词等, 如是, 则转有关分词典查找。以“加工”一词为例, 它在汉语词典上可以得到以下信息:

单词	词类	属	语义(组)	是否多义	是否成语	是否结构
加工	动词	一般动词	动作	是	否	否

4.2 成语词典

针对工艺文件中的一些超过 6 个字的较长的专有名词, 可以用专门的成语词典进行查找, 这种方法的优点在于迅速、准确而方便。例如: 机轮护板后接头, 滑油金属屑末信号器等。很明显, 作为成语处理的词, 如果被分开作为单词逐个加工, 结果是不理想的, 无论在速度或质量上都会受到影响, 经过成语词典的加工, 成语词在得到词义的同时, 还可得到它的语法功能信息。

4.3 同形词典

解决语法方面的多义是机器理解工艺语言不可避免的一个关键问题。多义词典的主要任务就是通过上下文分析的方法最终确定一个多义词在一个具体的语境中单一的、明确的语义和语法功能。例如: “加工”一词, 在下列两个例句中, 虽然词义相同, 但词类并不相同:

- 1) 加工零件外圆。
- 2) 完成外圆的加工。

前一句中为动词, 后一句中为名词, 为此建立同形词典及相关规则库。

4.4 结构词典

工艺语言词汇丰富, 表达方式变化灵活。有一些词加上与其搭配的某些词, 往往有较为固定的结构, 而这种结构紧凑、层次分明的表达形式往往有着明确的含义和清楚的语法成分。例如, 在这样一句话中: 内外缘卡板与角材的间隙应保证。分析其主语部分, 共得到 3 个名词: “内外缘卡板”、“角材”和“间隙”。根据汉语语法规则, 可以确定在这 3 个名词中, 中心词是“间隙”。由介词“与”和助词“的”联结“内外缘卡板”和“角材”组成短语修饰“间隙”, 组成名词词组(NP)。因此, 可以把这样一个结构归纳出来, 即: NP 与 NP 的间隙。用专门的结构词典加工类似的结构式, 大大节省加工时间, 并且有助于机器正确理解原文。

4.5 双语词典

双语词典是汉语和英语单词的对照词典。

4.6 英语词典

英语词典是为英文的形态生成作准备的, 它给出了一个英文单词的不同词形, 如动词的过去式和进行式等。

5 系统自动翻译的基本过程

5.1 汉语切分

因为汉语是一种孤立语,曲折变化少,而且汉语句子的词与词之间没有空白间隔标志,所以首先要对汉语句子进行单词的自动切分.汉语自动切词,系统采用最大匹配查词法.即由句首从左向右一次扫描 6 个字(不足 6 个字的有几个算几个字),到词典中查词,查不到从右削去 1 个字,再查,直到查到为止.每查到一词后,再将后面所剩的字补足 6 字再查找,如此循环往复,则可将 1 个句子中的全部词汇查出.

采用一次扫描 6 个字是综合考虑翻译的质量和速度确定的.一次扫描的字太少,许多较长的单词查不出来,影响翻译质量;一次扫描的字太多,会影响翻译速度.经过多次试验比较,最后选定 6 个字.

5.2 词性判断

句子切分为单词之后,就可以从词典中得到单词的初始信息(词义、词类信息).如果存在歧义,就需要进行词性判断.判断词性采用语法和语义结合的判断原则.规则库包括以下内容:

5.2.1 相邻规则(双向规则,即反之亦然的规则)

- 左邻为助动词(如“能”、“会”等词)不能取名词;
- 左邻为副词或助动词“地”,不能取名词;
- 左邻为语助词、副词,取动词;
- 左邻为形容词,不能取量词;
- 左邻为助动词“的”或形容词,则名动同形词取名词;
- 右邻为语助词(如“着”、“了”、“过”),则名动同形词取动词.

5.2.2 远程规则

- 右侧有方位词,取介词(如“当”);
- 右侧有可以共现的词,取介词,否则取连词(如“和……相同”、“就……而言”);
- 右侧有右邻不为“的”的动词,名、动兼类介词取介词.

5.3 确定主动词

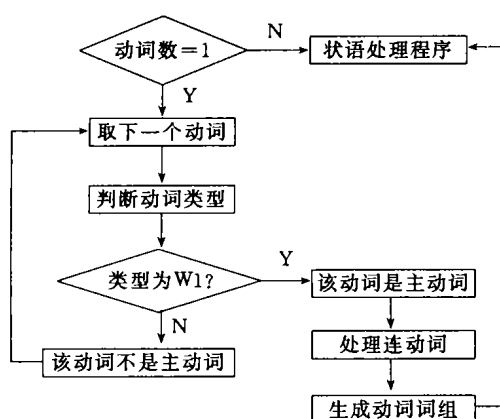


图2 确定主动词流程图

语法分析第一步是确定主动词.由于汉语的动词没有性、数、格和时态的形态变化,所以用语法确定哪一个是主动词是有困难的.一个句子的动词可能不止一个,那么确定哪个是主动词就是句子的核心.系统将一个句子中的动词按其分布情况分成三类.第一类是介词框架外的右邻不为“的”的动词(W1),第二类是在介词框架外的右邻为“的”的动词(W2),第三类是在介词框架内的动词(W3),只有W1类动词可以是主动词.所以,确定主动词有两个步骤,首先是对动词进行分类,将情况简化,然后再根据规则确定出主动词.确定主动词流程图如图2所示.

5.4 名词词组的合并

确定主动词后, 将词(主要是名词)合并, 达到同一语法块中的相邻词的词性互异. 在合成名词词组时, 要尽可能地考虑到有关名词的各个修饰语的各种组合的可能性. 因此, 系统把组成名词词组的各个转换规则有序地排列起来, 构成一个文法. 在工艺规程中, 系统考虑了各种常见语法形式的分析规则.

5.5 生成汉语完全语法树

完全语法树反映了可以涵盖任何工艺文件中的陈述句的深层结构: (定) + 主 + (状) + 谓 + (补) + (定) + 宾. 它可作为中间结果显示, 使句子的深层结构一目了然. 生成汉语完全语法树的基础是词性和动词类, 基本依据是属性制约短语文法. 在语法分析中就是要规定出句中不同功能块的边界. 在确定了主动词后, 以主动词为分水岭, 自左向右找出状语和主语, 向右扫描判断补语和宾语, 依靠属性制约文法就可以得到语法树.

5.6 其它成分处理

包括处理主动词前的主语和状语, 主动词后的宾语和补语.

工艺语言中出现主语的情况比较少, 一般都为动宾句. 综合工艺文件中的句子, 作者发现一种常见的主语和状语的歧义现象, 如下:

1. 接头是重要的零件.
2. 虎钳夹紧零件.
3. 孔口加工倒角.

以上3个句子中主动词之前都是一个名词. 但很显然, 第一句中的名词是主语, 后两个句子中的名词是省略介词的状语, 而且省略的介词也各不相同, 译文如下:

1. The connection is a special important part.
2. Tighten the parts with the cramping frame.
3. Machine the bevels by the porthole.

要消除这种歧义现象光靠语法分析是不够的, 需要辅以语义判断. 在汉语词典中给出“接头”、“虎钳”、“孔口”三个词的语义分别为“物体”、“工具”、“位置”, 制定相应语法和语义分析规则, 就可得出正确的译文.

如果主动词前为状语, 按状语判断规则进行分析.

宾语基本上就是前面合并的名词词组, 所以不再赘述.

5.7 中英文词汇转换

词汇转换的任务是把汉语分析阶段所产生的语法树上的词汇, 替换成英语的词汇, 并结合给出这些英语词汇的语法语义特征. 工艺语言汉英自动翻译系统的词汇转换工作, 可由一部双语词典完成. 越是深层的语言结构, 在不同语种之间, 它的差别越小. 所以在翻译时, 对原语的深层结构的分析结果, 可以少作或不作更动, 但是原语表层结果却必须更改. 为了作这种更动, 在实现词汇转换的词典中要提供出足够的表层结构信息. 只有这样, 在生成阶段才能根据这些信息, 生出符合英语语法要求的句子来. 词汇转换的输入和输出, 都是一个树, 其结构和语法分析结果相同, 只是叶上的汉语词, 被换成英语词, 各词语法语义特征, 也相应作了更换.

5.8 中英文结构转换

结构转换的主要任务是抹去汉语中对于英语来说多余的成分. 如: 汉语中的结构助词(的、地、得)、动词后助词(着、了、过)、动词前助词(把)和时间副词(已经、将)、量词等等. 把汉语中

的功能词转换为英语的相应形式,添加目标语言所必须的某些新成分.

5.9 英文生成

英文生成阶段的工作有两方面内容.一方面是从结构转换的语法树出发,建立译语的表层树;另一方面是把这个树形结构转换成译语的链,并根据上下文要求,生成每个词的词形.根据这两方面的不同要求,可以把生成阶段分为语法生成和形态生成两个子阶段.语法生成是对结构转换后的语法树,再进行转换,以生成能表示目标语言语法结构的语法树,该树形图的各末端结点上,要给出生成输出文句所必需的信息,并表示出输出句子的最后的表层结构的词序.形态生成是自动翻译的最后一个阶段,它要将目标语言的树形图变为语言文字链输出,至此,汉语句子的英语翻译就全部完成了.

6 后记

在进行理论研究的基础上,应用上述理论与算法,结合成都飞机公司的实际情况,开发了适用于成飞公司超七飞机制造工艺的翻译系统,该系统是基于工艺语言的以语法分析为主,并辅以逻辑语义分析的自动翻译系统.验证了前面所述的工艺语言汉英自动翻译系统的框架与结构基本合理,采用的原语自动分析和原译语自动转换以及译语自动生成的方法基本可行.汉英自动翻译系统最后与工艺文件编制系统(CAPP)实现了集成,生成了飞机工艺文件编制与汉英自动翻译系统.系统在生成中文工艺文件后,调用自动翻译系统进行翻译,得到英文工艺文件,并进行人工编辑,最后输出.

[参考文献]

- [1] 吴蔚天,罗建林.汉语计算语言学[M].北京:电子工业出版社,1994
- [2] 吴蔚天.汉语计算语义学[M].北京:电子工业出版社,1999

The Research and Practice in Automatic Translation for Machinery Technology

Zhou Jiru

(College of Electrical and Electronic Engineering, Nanjing Normal University, Nanjing, 210042, PRC)

Abstract: Based on the characteristic of language of technology, the researches are made on some problems concerned with Chinese-English automatic translation system for technological language. Combined with the theories expounded above, the CAPP and Chinese-English automatic translation system of Chengdu Aircraft Company are introduced.

Key words: manufacturing industry on a global scale; automatic translation; language of technology

[责任编辑:刘健]