

机器翻译模板的建立原则和方法

郑保山 刘群 张祥

中科院计算所 100080

摘要: 本文说明了模板技术在当前机器翻译系统中的重要性。在人工翻译和机器翻译经验的基础上,分析了模板的建立原则和方法,介绍了几种模板的生成、特点、消歧和用途,以及在英语词典出版用机器翻译系统中的应用。

关键词: 机器翻译 机译模板 机译词典 英语词典

Establishing Principle and Method of the Templates for Machine Translation

Zheng Baoshan Liu Qun Zhang Xiang

Computing Technology Institute of Chinese Academy of Science

Abstract: This paper points out the importance of a template technology in current machine translation system. On the basis of the experience of the artificial translation and machine translation, it presents the principle and method of establishing the templates. It also introduces the generation, characteristics, ambiguity resolution and use of some kinds of templates, as well as the application in machine translation system for publishing of English dictionaries.

Keywords: Machine Translation, Template of MT, Dictionary for MT, English Dictionary

1 引言

目前,机器翻译的研究正在与大规模语料库结合起来,加强对源语的研究,特别重视双语语料库和例句的作用,用以提高系统的适用性和译文质量。在新的系统研究和设计中,机译系统的使用范围正在缩小,主张面向一个特定的领域,面向一个具体使用对象的解决方案。在建造机器翻译系统的具体方法上,很多人觉得使用机器翻译模板是一个比较好的方法,可以显著提高系统译文生成的质量。^{[1][2]}

作者在与翻译出版界接触中,发现出版社最忌使用机器翻译系统,惧怕因此而把稿件搞得一团糟。作者也的确尝试过用机器翻译去翻译英语词典,不但没有省事,还徒增了许多麻烦。作者认为,对于像英语词典一类的源语,有现成的可以机器加工的带标文件,内容的重复率又比较高,其文本格式又非常适合于机器处理,翻译的劳动量又大,应当是机器翻译介入的领域。于是,决定专门从事英语词典翻译出版用英汉机器翻译系统的研究工作。^[3]

模板的思想在不少行业中造就了很多像样的产品，尤其是在软件中得到了广泛的应用。无以规矩不能成方圆。作者总结了使用机器翻译的方法翻译英语词典不成功的各个环节，全面细致地对英语词典语料进行了处理和分析，开始研究如何用机器翻译模板与人工翻译相结合的方法达到译文的比较精确的生成，从根本上减轻翻译英语词典中的繁重劳动，加快英语词典翻译出版的速度。

2 机译模板建立的原则和方法

2.1 模板的建立原则

所谓机译模板，实际上是一个从源语句子中抽象出来的句子标准框架，中间有若干个槽（slots），可以分别插入不同的词或词组（句片），因而成为一类句子的样板。这种模板建立的若干原则是：

（1）为了对付源语变化的随机性和开放性，框架的构筑必须使用可以穷举的词来充当。源语的实词是开放的，虚词如介词、连词、副词等是收敛的。因此，不能以实词为框架，必须以确定的词类——虚词为模板的框架。

（2）常用虚词差不多都有兼类现象。如果兼类中都是虚词，词典标注何类虚词都不关紧要；如果兼类中有实词，词典中不能标注实词，只能标注其中一种虚词类别。

（3）句法分析是以中心词（动词）展开的，中心词必须进入框架结构，包括动词的各种时态、语态和非谓语动词形式。

（4）模板可以做得简单些，也可以做得复杂些。简单的抽象度高，如全部用词类做成框架： $\#P[1]VP-\#P-\#V[2]\#P-\#V[3]ED-\#P[4]$ ，抽象度最高，因而覆盖面也最大。但是，它不能照顾每个动词间的差别，因而精确性不好。一般，以一个动词为中心，辅以其他动词的各种形式，并以虚词为框架制成模板，其精确性最好。

（5）常用动词的兼类现象很多，如名动兼类、名形动兼类等，都应在词典中标以动词词性，以免模板提取时发生遗漏。

（6）对于复合句，可以拆分成二个以上的模板，有利于简化模板的构成，提高通用性。

（7）只有一种模板，不能完成机器翻译的操作，至少要有源语模板和译语模板。实际上译语模板是源语模板的映射。为了插槽的方便，还要有插槽用的模板；为了统计的需要，最好还有抽象度最高的句型统计模板。

（8）如果采用插槽模板，槽内必须有指针和语义参数等，以进行槽内的各项操作，例如兼类的消歧。

（9）用于插槽的插板件——词或词组，最好以句片的形式进入机译词典。句片的多少对模板的简化和消除歧义有很大的影响。

（10）每句的源语模板是每译必生成的，基于模板的机器翻译系统是由源语模板驱动的。译语模板主要是人工制作的，需要精确生成译文的句子，一定要有相应的译语模板。对于某一领域句型统计结果为低频者（只出现一次），没有必要制成译语模板，代码的重用率太低，不如系统自动降为机助人译方式加以修改为好。

(11) 不论是源语模板，还是译语模板，它的初步形式都是自动生成的。作为一块模板，要想达到所需要的精确程度，必须经过仔细评估和修改才能投入使用。译语模板的正确制作，直接关系到译文的质量。

(12) 不是所有的句子都有模板可用的。译语模板达到的覆盖率，也就是系统译文质量的保证，如达到 60% 以上的覆盖率，整个系统的译准率肯定在其之上，它是完全可以重复的。对于没有译语模板的句子，系统采用自学习得来的例句翻译或传统机器翻译方法加以处理。

2. 2 模板的建立方法

(1) 一个应用领域确定之后，首先要收集这个领域有代表性的语料，建立这一领域的足够规模的语料库，并且需要不断地加以补充。

(2) 在词频统计的基础上，建立机译电子词典。词典中动词的频率参数，是建立模板的最好参考。

(3) 对应用领域的大规模语料进行句型统计，建立句型语料库。其中，句型的出现率是编制译语模板的重要参数。

(4) 通过句型语料库，再建立句片语料库，建立词组（句片）电子词典。

(5) 通过句子分析，建立源语句子模板、源语插件模板和译语生成模板。

(6) 源语句子模板用于源语句子的查询，如源语句子：

Scanners usually capture images at a resolution of 300 dots per inch (dpi).

源语句子模板为：usually-capture[2]at[3]of[4]per[5]([6])

(7) 源语插件模板用于插槽，如（6）的插件模板为：

[1.1,1]usually-capture[2,1,4]at[3,2,6]of[4,2,9]per[5,4,12]([6,1,17]).

注：[]内的数字分别为槽参数槽的序号、槽内的词（组）的个数和槽的指针。

(8) 译语生成模板用于译语的精确生成，如上述句子的译语模板为：

[1]通常以[5]的[4]的[3]捕获[2]([6])。

机器翻译结果为：扫描仪通常以每英寸 300 点的分辨率(dpi)捕获图像。

3 机译模板的消歧处理

机器翻译模板的构成特点是动词和虚词在框架上，其余实词在槽里。在框架上的词（组）制作译语模板时可以任意移动，而在槽里的词的词序是不能改变的。可能发生歧义的地方（该出现在框架上没有出现在框架上，该在槽里没在槽里）有如下几点：

(1) 由于对兼类词的词性标注，对实词偏向动词标注，对虚实词兼类者偏向虚词标注，因而它作为非标注词性出现的地方，就会干扰模板框架的形成。

(2) 形容词作名词的后置定语，如：available, possible 等，需要在槽内调序。

(3) 时间副词，如：this year, last Sunday 等，译语词序的变动更大。

(4) 数词与名词混在一起，如：The moon is about one-quarter the size of the earth..

(5) 普通疑问句，Will you do me a favour? 间接宾语与直接宾语同处一槽。

3. 1 兼类歧义

名动兼类在句中作名词而在框架中以动词出现，如：

We calculated keyboarding costs on the basis of 5,500 keystrokes per hour.

[1]calculated[2]on[3]of[4]keystroke-per[5]

在槽内作规则，改写 keystroke 的词性为名词，

则源语模板变为：calculated[2]on[3]of[4]per[5]

译语模板为：在每[5][4]的[3]上计算了[2]。

译语生成：我们在每小时 5,500 击键的基础上计算了键入成本。

3. 2 动词一词多义处理

动词一词多义是常见的现象，因为模板是带着语境而给出汉义，因而比较容易处理。例如，在计算词典机器翻译中，动词 allow 的出现频率是非常高的，需要形成若干块模板，可以给出比较贴切的汉义来。如：

The cable adapter allows attachment of the scanner to the SCSI interface.

[1]allow[2]of[3]to[4] [1]允许[3]的[2]接到[4]上。

电缆接合器允许扫描仪的附件接到 SCSI 接口上。

The software allows captured images to be edited.

[1]allow[2]to-be-edited [1]可以编辑[2]。 软件可以编辑抓取的图像。

3. 3 源语与译语用词遣句差别大

一般，槽中的汉义是查词典的直接结果，不能改。为了保证译语模板的精确性，根据行文的需要，可以在槽外给一个确切的含义，而将槽内的汉义括起来，作为备选。同时，对源语实施意译。如：

The car stopped short only a few inches from where I stood.

[1]stopped[2]only[3]from-where[4]stood

[1]停在这么近[2]的地方，离[4]站的地方只有[3]。

汽车停在这么近（短）的地方，离我站的地方只有几（不多）英寸。

3. 4 形容词比较级前的数词

形容词比较级与前面的数词作状语成分在同一槽中，必须将比较级形容词分出来，以便调整词序。如：He is two inches shorter than I. [1]is[2]than[3]

在槽[2]中含有 two inches shorter，只能译成“二英尺短”。

从槽[2]中分离出 shorter 后，源语模板自动生成：[1]is[2]shorter-than[3]

这样一来，译语模板可以作成：[1]比[3]矮[2]。他比我矮二英寸。

3. 5 数词与名词间加量词

数词后的量词，根据其后的名词语义参数而定，这些都需要在槽内解决。如：

There are two poems on the blackboard.	在黑板上有二首诗。
There are four electric lamps in the hall.	在大厅中有四盏电灯。
There are two tankers here.	这儿有二艘油船。
There are two blankets on that bed.	在床上有二条毯子。
There are ten rifles on the ground.	在地面上有十杆步枪。
There are two mirrors on the wall.	在墙上有二面镜子。

此外，冠词之译与否，是一个比较复杂的问题。同理，也可以在槽内通过语义参数的正确标注来解决。如：He is a senior engineer..（他是一位高级工程师。），Here is a telegram for you.（这是你的电报。）

4 模板在英汉机器翻译系统的应用

（1）由出版社提供的带标文件，经过数据清洁处理，生成源语语料库。从语料库统计出源语使用的词、词组、句片和句子，以及相应的使用频率。

（2）由词和词组的统计中抽出动词（包括 ING 和 ED 形式）、虚词（介词、连词和副词），建造模板生成用的框架词典。框架词典为英汉对照形式，标注有词类、语义参数和出现频率，并且在使用中统计调用频率。

（3）由词、词组和句片统计中抽出除上述动词和虚词以外的实词，建造模板生成用的槽内用词词典。词典结构与框架词典相同。

（4）对句子库做句型统计。采用上述以动词和虚词的词类为框架的句型模板，生成句型库。句型库的结构包括源语、句型和句频。

（5）对句频大于 1 的句型，生成模板库。模板库的结构包括源语、句型、源语模板、插槽模板和译语模板，另外还有调用频率一项。

（6）制作模板库中的译语模板项。译语模板项已经通过规则的运算自动填上译语模板的雏型，简化了制作过程。

（7）机器翻译主程序，除了一般的功能之外，每一个要翻译的句子都进行源语模板和插槽模板的标注，用源语模板去查模板库，查中后用插槽模板去填词取义。

（8）查不中译语模板的句子，系统提供三种办法处理。一是选择主程序的制作译语模板选项，可以就地制作译语模板；二是选择主程序的机助人译平台选项，机器翻译系统自动降为机助人译平台，由用户直接翻译，翻译结果系统自动学习，永久有效；三是按系统自动生成的不十分精确的译语模板直接输出机器翻译结果。

5 结束语

在长期的机器翻译使用过程中，无论是外译中还是中译外，都给使用者以极大的方便，减轻了劳动强度，提高了译文的水平。在市场经济下，人们对产品的要求在不断地提高，机器翻译作为一种产品面市，当然也不例外。对于使用机器翻译的人，有这样一个要求：对于常见的句子能不能翻译得像样一点，改过的句子能不能不再改，机器不能译的句子能否让用户来译。如果能这样，机器翻译就有了效率。

根据这种情况，本文提出了使用机译模板的译语精确生成的概念，并根据手头的任务做了一次实验。实验结果很令人欣慰：凡是做了译语模板的句子，翻译得都很通顺，没有词序和数据垃圾的问题。与其相配套，还增加了机器学习功能，正确翻译一次，永久受益。机器翻译需要反复翻译才能出结果，这一点尤其显得重要。

试验结果：词典的释义部分有 11,110 个句子，经过主副句的处理，生成 14,570 个子句。2 频以上的子句为 4819 个，占 33%；需要译语模板 1461 块，占子句总数 10%。也就是说，以 10%子句的模板，可以解决 33%子句的精确生成问题。对于英语 900 句，拆成 1244 个子句，2 频以上者为 297 个，占 23.9%；需要模板 62 个，占子句 5%。也就是说，以 5%模板，得到了 23.9%查中率。

机译模板究竟是怎么一个概念，大概还没有一个完全的定义。本文制作的模板，目的是使句子的分析始终不脱离句子，不脱离语境，不脱离上下文关系，更加便于理解。在模板内安排动词的语义、介词的搭配关系，副词的修饰和连词的用法，都是在修辞级上进行的，所以，译文的生成结果，非常漂亮：槽内词或词组或句片的插槽取义，通过规则使其灵活起来，全面提高了译文的精确性。模板制作多少合适？要由应用领域语料的统计结果而定。一般，起码要占句子总数的一半以上，越多越好。用户能得到这样一个结果，也就会满意的。源语模板是自动生成的，译语模板的生成是半自动的，比用规则调句型容易得多。

本文所说的精确生成，是相对的概念，与译文模板的制作水平直接相关。本文的研究工作刚刚开始，遇到的问题不少，方法也不够完善，还在努力。

参 考 文 献

- [1] 董振东. 机器翻译研究进展. 计算机世界, 技术专题 D2, 1998-4-13.
- [2] 王海峰等. 汉英双向机器翻译系统 BT863 的研究与实现. 情报学报, 1997, 16(5): 360~369.
- [3] 郑保山, 刘群, 张祥. 英汉机器翻译系统的建造 (用于英语词典翻译出版的专用系统). 机器翻译与计算机语言信息处理国际学术研讨会论文集, 1999-6-26~28, 452~457 页.