基于UAMRT的机器翻译方法*)

李玉鉴

(北京市多媒体与智能软件技术重点实验室,北京工业大学计算机学院 北京100022)

摘 要 本文提出了一种新的机器翻译方法,即基于UAMRT的机器翻译。该方法的基本思想非常简单:首先设计模板匹配替换通用算法UAMRT,然后利用UAMRT匹配句子中的源语言模板,并将其替换为相应的目标语言模板,从而实现对句子的翻译。在结合句型分析算法和从句分析算法的基础上,利用启发式搜索机制进一步提高了句子的翻译速度和质量。速度测试表明用该方法实现的英汉翻译系统在P-IV1.7G的计算机上翻译速度每秒可以达到1300个单词左右;质量测试表明该系统的性能在开发过程中仅仅通过增加更多的模板就会变得越来越好,而且在应用时与几种商用系统相比可以达到中等水平。

关键词 模板匹配替换通用算法,句型分析算法,从句分析算法,启发式搜索,机器翻译

Machine Translation Based on UAMRT

LI Yu-Jian

(Beijing Municipal Key Laboratory for Multimedia and Intelligent Software Technology, College of Computer Science and Technology, Beijing University of Technology, Beijing 100022)

Abstract This paper proposes a new approach to machine translation based on UAMRT. The basic idea of this approach is very simple: it is to design UAMRT, namely, a universal algorithm of matching and replacing templates, and translate sentence by using UAMRT to match source templates and replace them with their corresponding target templates. Combined with sentential form parsing algorithm and a clause parsing algorithm, a source sentence can be translated by heuristic scans into a number of candidate target sentences in a few translation hierarchies of classes composed of ordered source-target rules. The speed test shows that the English-Chinese translation system based on UAMRT can reach a translation speed of about 1300 words persecond on P-IV1.7G computer; and three quality tests show that the system can not only become better and better by adding more templates without changing itself in the development, but also reach a middle accuracy in the application when comparing with commercial competing systems.

Keywords Universal algorithm of matching and replacing templates, Sentence form parsing algorithm, Clause parsing algorithm, Heuristic search, Machine translaion

1 引言

机器翻译的方法大致可分为基于规则[1,2]、基于语料统计[3,4]和基于例子[5]这三种。基于规则的翻译系统在句法分析的过程中,可能会出现很多歧义结构,如果没有很好的算法,分析效率就会比较低;对于自然语言中不符合语法规则的句子,基于规则的翻译系统通常也难于给出正确的翻译结果。基于语料统计的翻译系统一般翻译速度比较慢,并且事先需要积累大规模的双语对齐语料库对系统进行训练。基于例子的翻译系统需要充分利用短语一级的实例碎片,但是利用的实例碎片越小,碎片的边界越难于确定,歧义的情况也越多,从而导致翻译质量下降。

本文将提出一种设计和实现机器翻译系统的新方法,即基于UAMRT的机器翻译方法。该方法以模板匹配替换通用算法(Universal Algorithm of Matching and Replacing Templates,简称UAMRT)为基础,通过结合句型分析算法和从句分析算法建立启发式搜索机制,实现从源语言句子到目标语言句子的自动翻译。UAMRT的功能是从输入句子中匹配源语言的模板,并利用相应目标语言模板把匹配成功的部分翻译成目标语言。在这里模板是指由终结符和非终结符组成的任意字符串,终结符和非终结符是乔姆斯基的短语结构语法理论中的两个基本概念,本文把它们分别称为常项和变项。对英语句子"What do you think of it ?"来说,每一个单词what,do,you,think,of,it以及问号"?"都是常项,全部由常项构成的句子,称为纯常项句。而对带有两个变项的句型"What do NP think of NP ?"(NP表示名词或代词类型)来说,称为模板句型。一个模板句型可能不包含任何常项,例如,NP VERB (VERB表示动词类型)。如果考虑的是词组,那么不带变项的词组(比如a lot of)称为纯常项词组,带有至少一个变项的词组(比如give NP to NP)称为模板词组。一个模板词组也可能不包含任何常项,例如,VERB NOUN(NOUN表示名词类型)。模板是模板句型和模板词组的总称。为了概念的统一起见,本文有时也把纯常项词组和一些特殊的纯常项句子称为模板。当一个模板中的所有变项被其所能取到的常项值替代时所得到的纯常项序列称为该模板的一个实例。例如,give the book to me 就是give NP to NP的一个实例。

*)本文受到北京市教委项目Km200310005013和北京市优秀人才专项经费资助。李玉鉴 副教授,博士后,主要研究兴趣是机器翻译、神经网络和认知科学。

基于 UAMRT 的机器翻译系统不仅能够简单有效地处理那些不太符合语法规则的句子，而且使机器翻译系统的实现在一定程度上转换成为组织模板翻译规则的工作，从而能够较大地提高机器翻译系统的开发效率和翻译质量。速度测试表明以该方法为基础实现的英汉翻译系统在 P-IV1.7G 的计算机上的翻译速度每秒可以达到1300个单词左右；质量测试表明该系统的性能在开发过程中仅仅通过增加更多的模板就会变得越来越好，而且在应用时与几种商用系统相比可以达到中等水平。

2 基于 UAMRT 的机器翻译方法及系统结构

从形式化的角度看，模板可以看作由若干常项和变项构成的任意具有一定稳定性的语言认知结构。一个源语言的模板可以表示为下面的符号串结构：

$$a_1a_2\cdots a_{i_1}V_1a_{i_1+1}\cdots a_{i_2}V_2\cdots\cdots V_ma_{i_m+1}\cdots a_{i_{m+1}} \quad \text{(str1)}$$

相应的目标语言模板可以类似地表示为

$$b_1b_2\cdots b_{j_1}U_1b_{j_1+1}\cdots b_{j_2}U_2\cdots\cdots U_nb_{j_n+1}\cdots b_{j_{n+1}} \quad \text{(str2)}$$

在符号串 str1 和 str2中，有下面的约束：

(1)$1\leqslant i_1<i_2<\cdots<i_{m+1}$，$a_1,a_2,\cdots,a_{i_{m+1}}$都是源语言中的常项，$V_1,V_2,\cdots,V_m$ 都是源语言中的变项，它们可以相同，也可以不同；

(2)$1\leqslant j_1<j_2<\cdots<j_{n+1}$，$b_1,b_2,\cdots,b_{j_{n+1}}$都是目标语言中的常项，$U_1,U_2,\cdots,U_n$ 是目标语言中的变项，它们可以相同，也可以不同；

(3)$U_1,U_2,\cdots,U_n$ 一般来说是 $V_1,V_2,\cdots,V_m$ 中所出现的变项，但在一些特殊情况下也可以是与 $V_1,V_2,\cdots,V_m$ 中所出现的变项有某种确定性转换关系的变项，m 和 n 可能不相等。

由于一个源语言模板可能与多个目标语言模板相对应，因此在模板对应规则库中，模板对应规则在一般情况下应表示为：

源语言模板→源语言模板类型：目标语言模板1|目标语言模板2|…|目标语言模板 n；

其中源语言模板类型是源语言中的变项；如果目标语言模板只有一个，$n=1$，否则 $n>1$。一个模板对应规则的例子是“give NP to NP→VI：把 NP 给 NP；”。

在基于 UAMRT 的机器翻译系统中，模板对应规则库的建立具有中心地位，但目前模板对应规则的获取和增加还只能通过手工完成，因此是一项非常艰苦和费时的工作。如果把待翻译的源语言句子 S 表示为词序列 $w_1,w_2,\cdots w_L$，模板匹配替换通用算法要解决的基本问题就是自动从模板对应规则库找到在 $w_1w_2\cdots w_L$ 中所出现的源语言模板，并将其按照对应的目标语言模板替换为目标语言中的词序列，这一过程称为模板匹配替换运算。由于自然语言的复杂性，相似或相近的模板对应规则可能会相互冲突。为了减少模板冲突，提高翻译速度和质量，对于基于 UAMRT 的机器翻译系统来说，还需要设计并实现一个句型分析算法和一个从句分析算法，句型分析算法用来判定输入的英文句子是特殊疑问句、一般疑问句、反意疑问句、陈述句还是感叹句等句型，从句用来判定输入的英文句子所包含的从句类型。在句型分析算法和从句分析算法的基础上，就可以在模板对应规则按照一定层次结构分别组织在多个不同的规则库文件中，从而把建立模板对应规则库的任务主要限制在增加各种短语和简单句型的情况，在处理比较复杂的句子时则可以利用模板的嵌套结构进行翻译。目前，本文实现的翻译系统大约包含3000条规则，这些规则主要根据下面2条组织原则按照一定的层次结构被分别存储在多个不同的规则库文件中：

(1)层次相同的模板对应规则尽可能放在同一个规则库文件中，结构类似的模板对应规则尽可能放在同一个规则库文件中，句型相同和从句类型相同的模板对应规则也尽可能放在同一个规则库文件中。

(2)在同一个规则库文件中，所有起始项相同的规则按照项数多者在前的原则依相邻顺序组织在一起。

在对输入的英文句子进行翻译时，基于 UAMRT 的机器翻译系统首先根据句子的类型和结构特征，利用句型分析算法和从句分析算法，判定它是特殊疑问句、一般疑问句、反意疑问句、陈述句还是感叹句等句型，以及它是否可能带有定语从句、宾语从句和状语从句等从句，同时记录句型或从句的关键位置标记；然后选择与输入句子可能相关的所有规则库文件，从而在规则库层次上限制可能参与模板匹配替换运算的规则总数。在与源语言句子进行模板匹配替换运算时，根据句型或从句的关键位置标记，按照先在可能相关的低层次规则库文件中搜索，后在可能相关的高层次规则库文件中搜索的顺序，寻找可能与输入句子完全匹配的模板对应规则。由于在同一个规则库文件中，所有起始项相同的规则按照项数多者在前的原则依相邻顺序组织在一起，因此在选定规则库文件后，还可以按照下面的启发式搜索机制，迅速确定输入的英文句子是否包含该文件中所具有的模板：

(1)将规则库文件中的起始项相同的规则分为一组，根据起始项确定各组的第一条规则和最后一条规则的序号(在实际的翻译系统中，这两个序号是系统启动时通过一个专门的算法自动获得的，它们可能随着规则的增加或减少而变化)。

(2)记录英文句子中的当前匹配位置，据此可以得到该位置所指向的当前匹配项。

(3)利用当前匹配项的单词原型和可能的类型，以及(1)中得到的序号数据，在规则库文件中确定可能与英文句子在当前位置匹配的规则子集。

(4)从当前位置开始，将英文句子中的各项与规则子集中的每个规则按顺序一一进行比较，如果找到在源语言模板部分完全匹配的规则，则退出比较，并根据对应的模标语言模板进行替换运算，得到一部分翻译结果。

不妨假设某个规则库文件中包含以下模板对应规则：

change to NP → vi. 改变成 NP；
change NP for NP → vi. 把 NP 换成 NP；
change NP into NP → vi. 把 NP 转变成 NP；
ask NP for NP → vi. 问 NP 要 NP；
ask NP ADV → vi. ADV 问 NP；
ask NP NP → vi. 问 NP NP；
develop from NP → vi. 从 NP 发展而来；
develop into NP → vi. 演变成 NP；
……

容易确定以 ask 为起始项的规则组中第一条规则和最后一条规则的序号分别是4和6，如果输入的英文句子是：“He asked me for money.”，那么在当前匹配项为“asked”时，可以在规则库文件中确定与输入句子匹配的可能规则集为第4～6条规则，直接将英文句子与第4～6条规则按顺序一一进行匹配运算，不难发现“He asked me for money.”与第4条规则匹配成功，所以被翻译成“他问我要钱。”

严格说来，一个完整的基于 UAMRT 的机器翻译系统应该包括6个主要模块，即：基本单词库，固定搭配连接词组库，模板对应规则库，词法分析算法，句型分析和从句分析算法，模板匹配替换通用算法。整个翻译系统的总体框架可以用图1来表示。

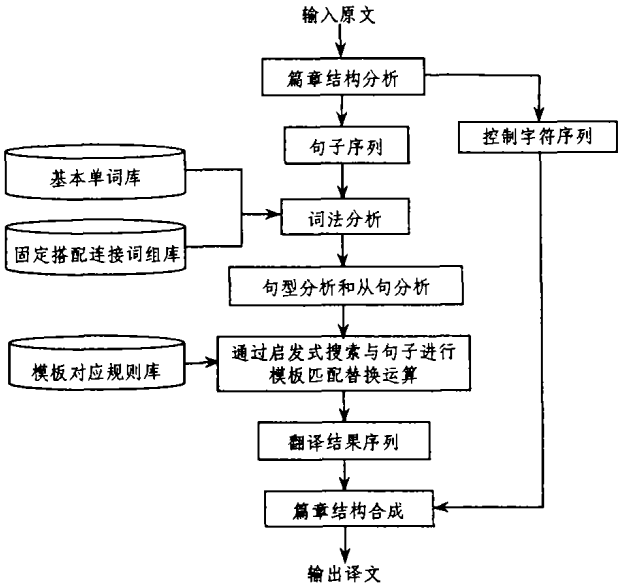

图1 基于 UAMRT 的翻译系统总体框架

3 系统的速度和质量测试

以基于 UAMRT 的机器翻译方法及系统结构为基础，本文实现了一个英汉翻译系统。该系统在 P-IV1.7G 的计算机上，翻译速度可以达到每秒1300个单词左右，而且这一速度不易受模板对应规则数目的影响，具有较强的稳定性。虽然这一速度并非最快，但是它已达到了市场上大多数同类系统的水平，而且在同样的计算机上要明显快于 IBM 翻译家2000的翻译速度。

为了能够比较客观地评价系统的翻译质量，本文共选择了三组测试数据。第1组数据是北京大学计算语言所提供的英汉翻译测试大纲中的470个具有代表性的英语句子（见 http://www.icl.pku.edu.cn/ research/mte/mte.htm），句子的平均长度是8.1个单词，进行封闭测试的结果表明，系统的翻译正确率可以达到93.6%；第2组数据是从《走遍美国》中抽取的1000个句子，句子的平均长度15.0个单词，进行封闭测试的结果表明，系统的翻译正确率可以达到63.1%；第三组数据是从多种英语学习资料中所抽取的1010个句子，进行开放测试的结果表明，系统的翻译正确率可达52.3%。其它用于比较的系统包括金山快译2000、东方快车3000、桑夏译王 Light 1.5.2，IBM 翻译家2000，所有测试列于表1中。

表1 几个典型英汉翻译系统的正确率测试结果

英汉翻译系统	金山快译2000	东方快车3000	Light 1.5.2	IBM 翻译家2000	本文的系统
封闭测试1	74.9%	88.7%	90.6%	77.9%	93.6%
封闭测试2	49.4%	58.0%	60.3%	55.8%	63.1%
开放测试	43.7%	55.0%	59.4%	45.0	52.3%

在进行以上测试过程中，一个翻译结果正确与否的判别标准是主要意思正确、词序基本通顺。从测试结果不难看出，与市场上流行的其它几个同类系统相比，本文的英汉翻译系统在正确率上已达到了中等水平，这说明它具有较好的发展潜力，因为它所使用的模板数量还相当有限（大约3000左右），而且在进行模板匹配替换过程中所使用的句型分析算法和从句分析算法还有待于进一步的改进和提高。

4 方法的特点和困难

4.1 模板的中心地位

本文的最大特点在于把模板的概念提高到了机器翻译系统的中心地位。模板的概念并非本文首创，它在机器翻译的研究过程中已经被或多或少地使用。比如，在文[2]提出的多策略机器翻译系统中，仅在模式库中存在与当前语句完全匹配的模式时，直接从事例模式库中获取相应模式的解模式，然后通过译文生成模块得到句子的译文。

例如，假如模式库中有一个如下的模式：

R (HV)is a AP(1)NP(1,HV)→{S,(),(!R 是一个!AP !NP。),()}

其中，R、AP、NP 分别代表语法类代词、形容词和名词的标识符；HV 表示人类的特征；1表示相应结构成分的归约长度为1；{S,(),(! R 是一个! AP! NP。),()}是解模式。对于与该模式完全匹配的句子"He is a good teacher."的译文生成模式为"!R 是一个!AP !NP。"，最终译文是："他是一个好的老师。"

上述例子说明，文[2]中的模式概念与本文的模板概念有一定相似之处，但不同之处在于，文[2]中的模式似乎仅指具有几个变项的简单句式，而本文的模板具有更广义的含义，可以指由若干常项和变项组成的具有一定稳定性的任意语言认知结构。更重要的区别在于，模式在文[2]中并不具有中心地位，对于模式库中存在与当前输入句子有多个相似模式的情况，以及不存在相同或相近模式的情况，模式并不直接用来构造翻译结果；而在本文中，模板是整个翻译系统的核心，对于任何输入的源语言句子，本文都首先进行句型分析和从句分析，然后借助精心设计的启发式搜索机制和模板匹配替换通用算法自动找出当前源语言句子中存在的模板嵌套结构，同时根据对应的目标语言模板自动构造最终翻译结果。由于模板对应规则库不够充分和完备的原因，本文所实现的英汉翻译系统暂且还不能对任意英语句子都给出正确的模板嵌套结构和令人满意的翻译结果，但是从所使用的方法来看却具有很好的领域扩展性，因此通过不断增加新的单词、词组和模板，翻译系统的性能就会逐步提高和完善。

4.2 与其它机器翻译方法的比较

基于 UAMRT 的机器翻译方法可以被看作基于规则的和基于例子的机器翻译的一种合理综合，因而在实现机器翻译系统时，比基于规则的和基于例子的方法具有更多的优势。在基于规则的方法中，通常难于从不太符合语法的句子中抽象出一条合理的规则，比如，考虑下面的句子，

It never gets as cold there as it does here.

如果采用基于规则的方法来翻译这个句子，就可能遇到比较大的困难。相反，如果采用基于 UAMRT 的方法来翻译，则只需给系统增加下面两条规则：

(1)get as ADJ ADV1 as NP AUX ADV2 → VI:ADV1 变得象 NP ADV2 一样 ADJDE；

(2)NP ADV VI →S: NP ADV VI；

具体的翻译过程是：

(1)将 get 匹配为 gets，cold 匹配为 ADJ，there 匹配为 ADV1，does 匹配为 AUX，here 匹配为 ADV2，将 gets as cold

there as it does here 合并为动词类型，并翻译为“在那里变得象它在这里一样冷”；(注意变项 ADJDE 的含义是自动去掉对应形容词变项 ADJ 在中文意义中末尾的“的”字)

(2)此时原来的句子已经变为 It never VI，模板匹配替换方法自动将 it 匹配为 NP，将 never 匹配为 ADV，从而实现 it never VI 与 句型 NP ADV VI 的正确匹配，最后得到翻译结果：

它从来不会在那里变得象它在这里一样冷 。

对同样这个句子，东方快车2000和金山快译3000的翻译结果分别是：

当它这里做，它从来不作为寒冷得到在那里。

不决变和它在这里做同样在那里冷。

此外，由于基于 UAMRT 的机器翻译系统能够把一个尽可能宽广的范围作为整体来进行分析，因此容易获得比基于规则的方法更好的翻译质量。

在基于例子的方法中，如何翻译一个完整的句子仍然是一个没有很好解决的问题。文[5]提出的方法需要建立一个具有良好信息组织结构的庞大的翻译例子库，即要求每一个翻译例子由三个部分组成：一个依赖于单词的源语言树(a source word-dependency tree ，简称 SWD)，一个依赖于单词的目标语言树(a target word-dependency tree，TWD)，和一个对应列表(a correspondence list，CL)。显然，建造这样一个翻译例子库也需要耗费很多人力、物力，因为除了对于结构比较规范的句子来说 SWD 和 TWD 可能被机器自动实现以外，要建立比较复杂句子的 SWD 和 TWD 以及要在 SWD 和 TWD 之间找到对应可翻译单元(translatable unit)的对应列表 CL 必须靠人工干预才能完成。相比之下，虽然基于 UAMRT 的方法需要手工建立模板对应规则，但是却比建立一个完整的翻译例子要容易得多。此外，在句型分析算法和从句分析算的支持下，可以把建立模板对应规则库的任务主要限制在增加各种短语和简单句型的情况，对于比较复杂的句子可以通过利用模板的嵌套结构进行翻译。这说明为了实现一个同样性能的翻译系统，只需建立更少的模板对应规则即可。

通过以上分析，不难看出，基于 UAMRT 的机器翻译合理地综合了基于规则和基于例子的翻译系统的优点，同时能够在一定程度上克服它们的缺点。

4.3 多译文的选择方法

在用模板匹配替换通用算法设计机器翻译系统时，可能遇到下面两种类型的歧义：

(1)多义词引起的歧义，例如 counter selling arts and crafts，虽然本文的方法能够根据上下文消除 counter 的兼类词词性歧义，即确定 counter 为名词，但是无法从 counter 的多个词义(计数器，计算器，计算者，柜台，筹码)中选取“柜台”之义。

(2)在翻译规则中源语言模板所对应的目标语言模板不唯一引起的歧义，例如：

VT NP1 with NP2 → VI：用 NP2 VT NP1 | VT 戴 NP2 的 NP1

这一翻译规则的两个实例分别是

see a boy with a telescope → 用一架望远镜看见一个男孩

see a boy with a cap → 看见戴一顶帽子的一个男孩

歧义的存在意味着同一个英语句子可能有多个不同的汉语译文。消解歧义的方法之一是给系统增加更多特殊的模板对应规则，方法之二是使用文[6]提到的 n 元统计模型，方法之三是将多个可能的翻译结果列举在对话框中用户去选择。本文的翻译系统目前采用最后一种方法，即由用户去选择或修改正确答案。事实上，本文的翻译系统设置了两种翻译模式：全自动翻译模式和机助翻译模式。在全自动翻译模式下，系统选择第一个翻译结果为最终翻译结果，本文第3部分的测试结果是在该模式下做出的；在机助翻译模式下，系统首先将输入的英语文本自动分拆成一组句子，然后一次一句地把每个句子的多个可能翻译列举在对话框，最后由用户从对话框中选择正确的译文作为最终翻译结果，如果用户找不到满意的译文，也可以直接参照英语原句进行人工修改或翻译。

4.4 主要困难

采用基于 UAMRT 的方法实现翻译系统，最主要的困难是建立模板对应规则库，也就是常说的知识获取问题。目前，在本文所实现的英汉翻译系统中，所有的模板对应规则都是靠手工编辑输入建立的，因而是一件非常费时费力的工作。虽然速度测试和质量测试已经表明利用手工编码的方式也可以利用基于 UAMRT 的方法实现一个达到初步商用水平的英汉翻译系统，但是仅靠手工编码难于给系统增加较大数量的模板，从而难于较大幅度地提高系统的性能。如何利用统计机器翻译的方法自动地从双语平行语料库中自动获取模板对应规则，将是本文进一步的工作重点。

结论 基于 UAMRT 的机器翻译方法，不仅能够合理地利用基于规则和基于例子的翻译系统的优点，而且能够在一定程度上克服它们的缺点。在实际应用中，该方法对于处理具有分离词组和不太符合句法结构的句子，比其它方法具有较明显的优势，同时它在一定程度上使机器翻译系统的实现转换成为按照一定的层次结构在若干个文本文件中组织模板对应规则的工作，因而能够较大地提高机器翻译系统的开发效率和翻译质量。速度测试和质量测试也进一步说明用该方法实现的英汉翻译系统在与几个商业化的英汉翻译系统比较时已经可以达到中等水平，如果继续改进句型分析算法和从句分析算法的性能，并给系统增加更多的模板对应规则，那么系统还有相当大的发展空间。

参 考 文 献

1 Chen Zhaoxiong. Functional Architecture of SC-Grammar. Journal of Computer (In Chinese), 1992, 15(11): 801～808

2 Huang Heyan, Chen Zhaoxiong , Song Jiping. The Design and Implementation Principle of an Interactive Hybrid Strategies Machine Translation System IHSMTS. Journal of Chinese Information Processing (In Chinese), 1999, 13(5): 43～50

3 Brown P F, et al. A statistical approach to machine translation. Computational Linguistics, 1990, 16(2): 79～85

4 Brown P F, et al. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 1993, 19(2): 263～311

5 Sato S. MBT2: a method for combining fragments of examples in example-based translation. Artificial Intelligence, 1995, 75: 31～49

6 KAKT S, Yamada S, Sumtia E. Scoring multiple translations using character N-gram. In: Proc. 5th Natural Language Processing Pacific Rim Symposium 1999 (November 5-7, Beijing), pp. 298～302