面向北京奥运会的定制化英汉机器翻译系统*

宋金平　肖健　孙广范

中国电子信息产业发展研究院　　北京 100040

E-mail: songjp@ccidit. com　xiaojian@ccidit.com　morgan2001_sun@163.com

摘　要：本文介绍了面向北京奥运会的定制化英汉机器翻译系统的技术原理及翻译流程，重点介绍了针对奥运会特殊需要所进行的翻译模板定制化：局部翻译模板定制化和整句翻译模板定制化。局部翻译模板定制化包括可变词典类、动词搭配类、名词搭配类的定制化；整句翻译模板定制化包括从简单到复杂的共三类翻译模板的定制化。经过上述定制化后，面向北京奥运会的定制化英汉机器翻译系统的译文质量有显著提高。

关键词：机器翻译，定制化，翻译模板

A Custom English-Chinese MT System for Beijing Olympics

Song Jinping,　Xiao Jian,　Sun Guangfan

China Center for Information Industry Development, Beijing, 100044

E-mail: songjp@ccidit. com　xiaojian@ccidit.com　morgan2001_sun@163.com

Abstract: This paper introduces the technical principle and flowchart of a custom English-Chinese MT System for Beijing Olympic Games, and describes the custom work of the translation template. The translation template customization includes the customization of local translation template and whole-sentence translation template. The local translation template comprises of variable dictionary, verb collocation and noun collocation; the whole-sentence translation template comprises of three types of translation templates. After the customization, the quality of the translation of the customized MT system has improved significantly.

Keywords: machine translation, customization, translate template

1　引言

目前，传统的、通用的机器翻译系统翻译质量还不是很高，但针对某个领域进行定制化开发后，翻译质量就会有很大提高，同时，其实用性也会提高很多。2008 年奥运会的赛前、赛时和赛后有大量的外文资料需要翻译，单靠人工翻译很难保质保量完成任务。为此，我们针对奥运会的特殊需求，定制化开发了一套奥运语言资源库系统，该系统最终成功地用在了北京奥组委语言翻译服务中，并取得了良好的效果，这也是奥运百年历史上首次使用机器翻译系统。英汉机器翻译系统是这个奥运语言资源库系统中最重要的一部分，它的定制化开发的是否成功，是整个奥运语言资源库系统能否成功的关键。

北京奥运会的语言服务与翻译环境支持是奥运会筹备阶段和运行阶段中的一个重要环节。面向北京奥运会的定制化英汉机器翻译系统用于解决数据资源再利用和数据共享问题，提高翻译人员的工作效率及翻译质量，满足奥运会期间场馆和总部密集和即时的翻译业务需要。同时在该项目中，基于汉英比较语言学和可比语料库的研究方法，提高了译文用词用语的规范化和本地化，有利于国际交流与沟通。

2　本定制化机器翻译系统翻译技术原理及流程

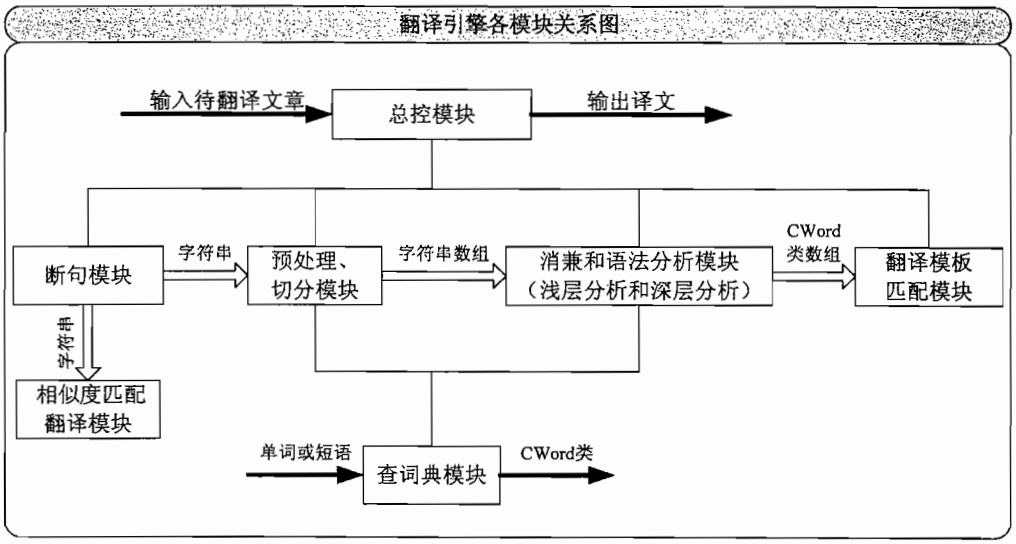
本定制化机英汉机器翻译系统采用混合策略的机器翻译引擎，是将基于规则方法、基于模板方法和基于实例方法相结合，将浅层分析方法和完全分析方法相结合的混合机器翻译策略。该

*本文受国家自然科学基金项目（项目号：60872118）资助。

方法是对我们的“基于语言知识库的机器翻译方法与装置”专利（专利号为：2004100011873）的活用和发展，流程图参见参考文献 [1]。

传统的规则机器翻译方法采用“完全语法分析+转换生产”模式，而这种模式往往导致完全语法分析质量不高，生成的语法树经常会出现错误，导致基于该语法树进行转换并生成的译文质量不高。本混合策略的机器翻译方法本质上是传统的规则机器翻译方法的延伸和发展，它采用一种基于“简单语法分析+整句翻译模板匹配”模式来代替传统的基于规则的方法中的“完全语法分析+转换生成”模式，从而可以基本解决长期困扰机器翻译系统的两个难点：即单纯基于实例的机器翻译系统在实际翻译过程中匹配成功率过低，大量句子无法成功翻译成译文，而单纯基于规则的机器翻译系统由于知识颗粒度太大，导致即便能翻译成译文，但译文可读性不高的问题。

本项目混合策略的机器翻译系统按功能可以划分为七大模块：



如上图所示，系统包含七大模块：

总控模块：控制翻译流程。

断句模块：对输入的文本进行断句处理。

相似度匹配翻译模块：对输入的句子和实例库中句子进行相似度匹配，匹配成功的直接输出译文。

查词典模块：用于查询词典的信息。

预处理、切分模块：对句子进行预处理，并把句子切成词和词组。

消兼和语法分析模块：用于分析句子的语法信息，包括切分校正、词性标注和合一运算 3 部分。

翻译模板匹配模块：对语法分析的结果形成的句法树与翻译模板库中的翻译模板比较，生成译文。

如上图，整个翻译流程由总控模块操控。程序的输入是一篇英语文章 A。一篇文章进入总控模块后，由总控模块调用断句模块，把文章断成一个一个的句子 A = {a1, a2, …, an} 。然后在预处理、切分模块中对单个句子进行切分处理，切分程序需要用到查词典模块，经过切分模块后，句子就变成了一个一个的单词或短语组成的序列：

A = {a1,a2,…,an}；其中：

a1 = u11u12…u1m1;

a2 = u21u22…u2m2;

……;

an = un1un2…unmn;

过了切分模块，就是消兼和语法分析模块，首先需要对每个单词调用查词典模块，读入词典中的各种属性信息，这期间，如果遇到未登录词，系统会进行智能判断，如数字，人名等，并且自动添上相应的属性。然后根据读入的属性，对句子做一个简单的分析，消除切分错误，确定每个单词的词性，并且对名词短语、动词短语等进行合一运算，在这个过程中需要用到短语库、片语库和词义选择库等对比语言资源库。下一步就是翻译模板匹配模块了，根据语法分析的结果，再在翻译模板库中找出与该语法分析结果匹配度最高的的翻译模板。方法是：对于输入的句子 ai，首先要计算它与翻译模板库中每个模板 cj : bj　ej 左边条件部分的匹配度 sij，这里匹配度可根据下表进行计算。

Object and logic Relation	Weight
“X”，“Y”，…	0
“SV”，“VO”，…	-1
AND NOT	-1
“VZhu”，“VBei”，“Ving”，“Vedo”，“Ved0”，“VV”，“NP”，“NONE”，…	-2
“A”，“N”，“V”，“I”，“J”…	-3
“SUB”，“OBJ”，“MAN”，“MAN1”，“MAN2”，“IO”，“I1”，“Vs”，“Ving” “Ved1”，“Ved2”，“_Ved1”，“_Ved2”，“Aer”，“Aest”，“Ns”，“Ny”，“Ay”，…	-5
“N[I]”	-15
“N[A]”，“V[A]”，“N[A|B]”，…	-5
Special word	-15
……	……

匹配度表

输入句子 ai 的序列为：

ai = ui1ui2…uimi;

翻译模板库中模板 j 的条件部分可以分解为：

bj = vj1vj2…vjmj;

这样，

$$sij = \left| \sum_{k=1,\ldots,\ \max(m_i, m_j)} w(uik, vjk) \right|$$

有了该句子与翻译模板库中每个模板的匹配度值，选取其中匹配度最高的模板作为翻译模板，并按照模板右部 dl 进行翻译。

$$l = \arg\max_{j=1,\ldots,m} \{sij\}$$

3　翻译模板定制化

针对奥组委的特殊类型的翻译文本，我们对本系统的翻译模板进行了定制化开发。本系统的核心是翻译模板，它可以分为两类，即局部翻译模板和整句翻译模板。模板实际上是一种词汇化的规则。一个单语的模板一般是一些常量和变量组成的序列。常量表示具体的词语（终结符），而变量表示一类词和短语（非终结符），一个翻译模板由两个双语模板及其变量的映射关系组成。

1)　局部翻译模板定制化

局部翻译模板用来处理一个句子中局部片断的翻译，它在英语分析和生成阶段都发挥重要作用。系统支持3种类型的局部翻译模板：

(1) 可变词典类

```
give ... a second chance to\V\使...能有第二次机会
   & CAT[V] M_SEM[A] E_VAL[I] CODE[L] $
with ... in mind\F\有鉴于...
   & CAT[F] M_SEM[I] FSORT[1] $
take ... under one's wing\V\把...招至旗下
   & CAT[V] M_SEM[B] Z[7] $
put into developing ... as a player\V\倾注心血培养...成为运动员
   & CAT[V] M_SEM[B] $
spend ... seasons with\V\为*效力...个赛季
   & CAT[V] M_SEM[A] E_VAL[1] CODE[U] $
```

可变词典类中的省略号可以代表句子中数量少于或等于5个单词的任何成分，如：

“take ... under one's wing --> 将...招至旗下”，可以将片断 take Mr. Wang under one's wing 翻译成“将王先生招至旗下”。

(2) 动词搭配类

```
{edge} {REN} {for} {OBJ} --> %0[勉强战胜]%1 获得%3
{outsourse} {OBJ} {to} {OBJ} --> %0[把]%1 外包给%3
{kick} {habit} {out of} {OBJ} --> %0[改掉]%3 中%1
{compete at} {OBJ} {for} {GUO} --> %0[代表]%3 参加%1
{pay} {OBJ} {in} {bail} --> %0[支付]%1 的%3[保释金]
```

这类模板中含有变量，如REN 代表“人”，OBJ代表对象等。

(3) 名词搭配类

```
{measure} {toward} {OBJ} --> 为%2 采取的%0[措施]
{leak} {of} {N} {to} {MAN2} --> %0[把]%2 泄露给%4
{replacement} {of} {N} {with} {N} --> 把%2%0[替换]成%4
{pressure} {against} {N} {to} {V} --> 要求%2%4 的%0
{debate} {among} {N} {over} {N} --> 在%2 中间针对%4 的%0[争论]
```

和上面一样，这类模板中含有变量，如REN 代表“人”，OBJ代表对象等。

针对北京奥组委的特殊需求，系统在原有8000多条局部翻译模板的基础上，又定制化开发了5000多条局部翻译模板，从而保证了系统局部翻译结果的流畅性。

2) 整句翻译模板定制化

句子是语言中具有完整意义的最重要的基本单位，解决了句子的翻译问题，就解决了机器翻译中的关键问题。因而我们在设计本机器翻译系统时，以句子及句型为基本考察与研究对象，这种句型在我们的系统中就是表现为整句翻译模板。

整句翻译模板用来处理整个句子的翻译。它在句子的分析完成之后才发挥作用。虽然整句翻译模板在语法分析之后进行，但它具有优先级最高的特点，表现在它可以校正语法分析模块中的错误。

本系统的整句翻译模板库分三类：

第一类句式为只有主、谓语，或只有主、谓、宾语的模板；

第二类句式为除了主、谓、宾语，还带有其他句子成分的简单模板，其他成分是指状语或虚词等；

第三类句式为上述两类模板以外的复杂模板。

模板匹配按照从三类句式到二类句式，再到一类句式的顺序进行匹配，也就是先把复杂句化成多个简单句，再翻译这多个简单句，再组成译文。

下面是整句翻译模板的典型例子。

```
No loitering .—>禁止游逛。
{SUB} {V} as early as {N[I]}-->%1 最早在%3%2
It be not clear what {VO} .-->不清楚是什么%1。
There be more {MAN} in {N[!I]} than in {N[!I]} .-->%2 里的%1 比%3 里多。
{SUB} , however , {VX} .-->然而，%1%2。
{X} , {SUB} said in an interview broadcast {N[I]}-->%2 在%3 播出的一次访谈中称，%1
" {X} " {SUB} be quoted as saying by {OBJ} on {N[I]} .-->%3%4 援引%2 的话说，“%1。”
{U&&YEAR[1]} for {GUO} , against {N} , {GUO}-->%1 年代表%2，在%4 参赛，对手是%3
```

针对北京奥组委的特殊需求，系统新定制化开发了 8000 多条整句翻译模板，从而保证了系统整句翻译结果的流畅性。下面是两个整句翻译模板定制化的例子。

例句一：

```
北京奥组委待翻译句：
1997 for Australia,against Germany,Pakistan.
定制化新整句翻译模板：
{U&&YEAR[1]} for {GUO} , against {N} , {GUO}-->%1 年代表%2，在%4 参赛，对手是%3
匹配上该模板后翻译结果：
1997 年代表澳大利亚，在巴基斯坦参赛，对手是德国。
```

例句二：

```
北京奥组委待翻译句：
“She was named Victoria's Most Improved Diver in 2001, ” he was quoted as saying
 by a series of papers.
最初翻译结果为：
“她被评为 2001 年的维多利亚进步最大的跳水运动员”，他援引通过一系列的文章说。
定制化新整句翻译模板：
" {X} " {SUB} be quoted as saying by {OBJ} on {N[I]} .-->%3%4 援引%2 的话说，“%1。”
后，翻译结果为：
一系列的文章援引他的话说，“她被评为 2001 年的维多利亚进步最大的跳水运动员。”
```

4 结束语

本英汉机器翻译引擎参加了 2005 年度全国 863 机器翻译评测，其 BLUE 值为：对话：0.3776，篇章：0.3709。

本英汉机器翻译引擎在 2008 年第四届全国机器翻译（CWMT2008）评测中，在英汉新闻主系统领域，取得了 BLEU 值第二，NIST 值第一和 WoodPecker 值第一的较好成绩（详见参考文献[5]），如下表所示：

CWMT2008 英汉新闻主系统评测结果

受限情况	单位	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
不受限	CCID	0.3157	0.2542	9.5048	9.5143	0.7754	0.6468	0.4048	0.3603

本文对基于混合策略的英汉机器翻译系统作了一个简单描述，并对定制化方法进行了一个简单的说明。本系统的定制化开发实践以及北京奥组委在使用后良好的反馈证明，良好的定制化开发能有效提高翻译质量和效率，减轻翻译人员的工作量。机器翻译系统质量的提高是一个长期

的过程，目前基于统计的机器翻译系统发展很快，在局部片段的翻译上，统计机器翻译系统有较大的优越性。本系统也借鉴了统计机器翻译的长处，在术语提取上采用了基于对数线性模型的统计方法。本系统的译文生成来自于模板，是一个相对直接的转换，这也是本系统的最大优点。

以上是对混合策略的英汉机器翻译系统所作的分析，旨在说明这种机器翻译方法的实用性与前瞻性。机器翻译系统的开发极易陷入纯研究纯规则或者头痛医头脚痛医脚的纯经验主义误区。我们采用的方法既强调实用性，避免脱离语言实际的抽象规则误区，又强调累积性和扩展性，通过不断积累术语库、局部模板库、整句模板库，词义选择库，片语库等，保证系统翻译质量能够稳步提高。

参 考 文 献

[1] 孙广范，宋金平，袁琦（2006）．“基于混合策略的汉英双向机器翻译系统的设计”．2005 年度 863 中文信息处理与智能人机接口技术评测研讨会论文. 中文信息学报 2006 年 3 月. 北京

[2] 孙广范，宋金平，袁琦（2006）．“机器翻译中规则和模板的协调方法研究”．2005 年度 863 中文信息处理与智能人机接口技术评测研讨会论文. 中文信息学报 2006 年 3 月. 北京

[3] 鲁孝贤 “机器翻译语义排岐的方法”中国科技翻译 2007 年第 4 期

[4] 刘洋 刘群 “机器翻译评测中的模糊匹配”第二届全国学生计算语言学研讨会论文集 2004 年 8 月

[5] 孙广范，宋金平，万缨，许亮，肖健. “第四届全国机器翻译研讨会 CCID 技术报告”．第四届全国机器翻译研讨会论文集. 2008 年 11 月. 北京

[6] 侯宏旭，刘群，张玉洁，井佐原均. “2005 年度 863 机器翻译评测方法研究与实施”.中文信息学报 2006 年 3 月. 北京

[7] 王惠，詹卫东，刘群，<<现代汉语语义词典>>的概要及设计，<<1998 中文信息处理国际会议论文集>>，清华大学出版社，1998 年 9 月

[8] 梁颖红，赵铁军，翟舒，规则与边界统计相结合的英语基本名词短语识别，<<语言计算与基于内容的文本处理>>，清华大学出版社，2003 年 7 月

[9] Nyberg and Mitamura (1992) "The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains" Proceedings of COLING-92

[10] Takeda, K., "Pattern-Based Context-Free Grammars for Machine Translation," Proc. of 34th ACL, pp. 144-151, June 1996.