基于文本-模板直接匹配的机器翻译系统 *

吴闯[1] 吴宏林[1] 张俐[1] 刘绍明[1,2]

1. 东北大学 信息学院 计算机软件研究所 自然语言处理实验室，沈阳 110004；

2. 日本富士施乐公司 日本神奈川 2590157

摘要： 模板匹配是影响基于模板的机器翻译（PBMT）性能的关键因素。本文提出了一种面向机器翻译的文本-模板直接匹配算法。该算法可绕过模板抽取步骤，将待翻译句子和实例库中的模板直接进行匹配，以避免复杂的语法分析。同时我们构建了基于文本-模板直接匹配的翻译引擎，并在引擎中引入模板选优模块解决模板误匹配问题。实验表明该方法可以达到令人满意的性能。

关键词： 基于模板的机器翻译，模板匹配，模板选优

Machine Translation Based on Direct Matching Between Text and Pattern

Wu Chuang[1], Wu Hong-lin[1], Zhang Li[1], Liu Shao-ming[1,2]

1. NLP Lab, Institute of Computer Software, Northeastern University, Shenyang, 110004, China;

2. Corporate Research Group, Fuji Xerox, Co., Ltd., Kanagawa 2590157 Japan

Abstract: Patten matching is the key factor which influences the performance of pattern based machine translation. The paper proposes a direct matching between text and pattern algorithm. It uses the direct matching between the given sentence and pattern to avoid the difficult linguistic analysis for extracting pattern. In addition, we set up a PBMT Engine based on this algorithm. And the Engine involves the pattern-selected to avoid the problems that come from the miss-matching of patterns. The experiment shows us a satisfactory result.

Key words: pattern based machine translation, pattern matching, pattern-selected

1 引言

PBMT 作为基于实例的机器翻译（EBMT）的一种典型翻译方法[1]，克服了基于规则的机器翻译系统(RBMT)具有的语言知识获取代价大，语法、语义分析困难的缺点。

模板匹配是影响 PBMT 性能的关键因素。传统的模板匹配方法先将待翻译句子通过句法分析抽取出模板[2-6]，然后计算抽取的模板和模板库中原语模板间的相似度。虽然从理论上讲句法信息可以为翻译提供更多的帮助（如解决词语的远距离依赖问题），但由于目前句法分析的性能还很难满足实际需要，基于句法分析的 PBMT 性能并不理想。

本文提出了基于文本-模板直接匹配的算法，首先将待翻译句子根据源语模板进行变换，以变换后句子中的片段（连续的词串）为基本单位，计算变换后句子和源语模板之间的编辑距离。该匹配算法绕过模板抽取，从而避免句法分析。此外，本文在基于文本-模板直接匹配的翻译引

基金项目： 辽宁省自然科学基金资助项目（20072032）；沈阳市科学技术资助计划（1081235-1-00）

作者简介： 吴闯（1985—），男，硕士生，主要研究方向为自然语言处理；吴宏林（1978—），男，讲师，博士，主要研究方向为自然语言处理；张俐（1961—），女，副教授，博士，主要研究方向为自然语言处理；刘绍明（1962—），男，主任研究员、教授，博士，主要研究方向为自然语言处理、图论、模式识别。

擎中引入分词修正模块解决由于分词粒度过细和组合型歧义造成的句子变换错误问题；引入模板选优模块解决模板误匹配问题。

2 基于文本-模板直接匹配算法的机器翻译引擎

本文将 PBMT 看作一系列映射的过程，其模型如图 1 所示：

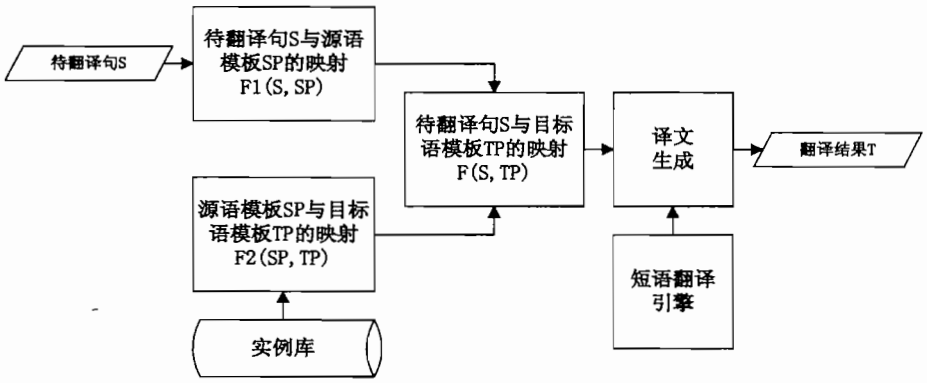

图 1 基于模板的机器翻译模型

对于待翻译的句子 S，首先建立 S 与源语模板 SP 的映射 F1(S,SP)。然后根据 SP 与相应的目标语模板 TP 的映射 F2(SP,TP)，建立 S 与 TP 的映射 F(S,TP)，在短语翻译引擎的支持下，生成译文 T。

文本-模板直接匹配算法中使用的模板结构如下：模板对 P 表示为 P=<SP,TP,Align>，其中 SP 和 TP 是互译的源语模板和目标语模板：

$$SP =< SP_1, SP_2, \cdots, SP_i, \cdots, SP_m > \tag{1}$$

$$TP =< TP_1, TP_2, \cdots, TP_j, \cdots, TP_n > \tag{2}$$

其中，SP_i 和 TP_j 分别为源语模板项和目标语模板项。模板项可以用一个二元组来表示：

$$PItem =< Item\_type, Item\_content > \tag{3}$$

其中，Item_type 是模板项的类型，Item_type∈{F,V}，如果一个模板项的 Item_type=V，称这个模板项为可变项，模板可变项主要为不影响结构的名词性短语。对于可变项，其 Item_content 为该可变项可以替代的文本类型，如：普通名词性短语、时间短语、处所短语等；如果一个模板项的 Item_type=F，称这个模板项为固定项。模板固定项主要为影响模板结构的动词、介词、助词、副词等，这些词一般是最为活跃的语法成分，翻译难度较高。这些词的变化，往往导致整个模板框架、以至于配价关系的变化，因此，将这些词作为模板固定项。对于固定项，其 Item_ content 为该固定项对应的文本内容。

模板对中的 Align 为源语模板和目标语模板之间的对齐关系。

$$\text{Align}=\{<i,j>|1\leqslant i\leqslant m,1\leqslant j\leqslant n\} \tag{4}$$

2.1 句子变换

设 S 为待翻译的句子，以词序列的形式表示：$S=<W_1,W_2,\cdots W_k,\cdots W_n>$。句子变换模块将 S 根据模板直接进行变换，定义变化后的句子为 $S'=<f_1,f_2\cdots f_k\cdots,f_n>$。其中 $1\leqslant k\leqslant n$，f_k 为一个变换后句子片段。变化后句子片段：f = <fra_type,fra_Info>，其中 fra_type∈{F,V}代表片段类型，若 fra_type= V 则表示可变片段，用 F_v 表示，否则表示固定片段，用 F_f 表示，fra_Info 表示片段内容。

2.1.1 句子基本变换

将 S 根据候选模板库中任意模板对的源语模板 SP 按照下面算法进行变换：对于任意一个模板固定项 SP_i：

⑴若 W_k 和 $Item\_content_i$ 相同，则生成固定片段 f_k，否则生成可变片段 f_k；

⑵若 f_k 是可变部分且 f_{k-1} 是可变片段，则合并 f_{k-1}，f_k 为一个可变片段。

2.1.2 分词错误修正变换

在句子基本变换中存在着由于分词粒度过小和组合型歧义造成的变换错误。如：分词粒度过小造成句子“将③与②上层的齿轮面嵌合”中的“嵌合”被分成两个独立的词，当“嵌”或“合”与模板“将 NP 与 NP 嵌合”中固定项“嵌合”比较时均不匹配，所以被当作可变片段与前面的“②上层的齿轮面”合并成为一个可变项，导致变换错误。本文引入分词错误修正模块，解决了基本变换中由这两类错误造成的变换错误问题。修正算法如下：

⑴把待翻译句 S 中与模板 SP 中的连续的 m 个固定项相同的单词 sk 修改为 m 个连续的词，分别作为 S'的固定片段；

⑵把待翻译句 S 中与模板 SP 的固定项 SP_i 相同的部分(连续单词串)sk 做为 S'的固定片段，同时把待翻译句 S 中的连续单词串修改为 1 个词；

⑶把待翻译句 S 不能固定的连续单词串分别作为 S'的可变片段。

2.2 变化后句子和原模板映射关系获取

编辑距离（Edit Distance）是由俄国科学家 Vladimir Levenshtein 于 1965 年提出的[7]，用于计算字符串的相似度。本文对变换后的句子 S'和源语模板 SP 采用编辑距离的方法计算相似度，用动态规划的算法求解编辑距离并生成编辑路径即变换后句子和原模板之间的映射关系 F(S',SP)。

2.3 次匹配判断及映射关系修正

对于任意的映射 F(S',SP)∈FSet(F',SP)，在 F(S',SP)中，如果 S'中有 1 个和源语模板 SP 的模板项没有对应关系的固定片段 f_i，且 f_i 满足下面条件时，则判定该 F(S',SP)需要修正：

⑴：S' 中有固定片段 $f_k(k\neq i)$，$f_i=f_k$ 且 $(f_k\leftrightarrow SP_j)\in f(S',A)$；

⑵：f_{i-1} 或者 f_{i+1} 为可变片段。

判定需要修正的映射为 F1(S',SP)∈FSet(S',SP)，同 SP 没有对应关系的 S'的固定片段为 f_i：

⑴：若 f_{i-1} 和 f_{i+1} 是可变片段，合并 f_{i-1}、f_i 和 f_{i+1} 为 S'的一个可变片段；

⑵：若只有 f_{i-1} 是可变片段，合并 f_{i-1} 和 f_i 为 S'的一个可变片段；

⑶：若只有 f_{i+1} 是可变片段，合并 f_i 和 f_{i+1} 为 S'的一个可变片段。

2.4 模板选优

通过句子变换模块，编辑距离计算模块和次匹配修正模块得到变换后句子和原模板集合 FSet(S',SP)，可能包含多个满足编辑距离阈值的候选模板。本文引入模版选优模块，获得

FSet(S',SP)中的最优匹配模板。影响模板选优的因素主要有三个，分别是S'和SP之间的编辑距离、短语译文翻译质量和候选模板覆盖质量。

变换后的句子和原模板之间的编辑距离反应了句子和模板之间的差异程度。如果S'和SP之间的编辑距离不为0，说明需要通过插入或者删除操作才能使S'和SP完全匹配，句子和模板之间较大的结构差异会影响最终的译文质量。本文定义句子和模板的编辑距离评价公式为：

$$DisScore(F(S',SP)) = |S'| + |SP| - Dis \tag{5}$$

短语译文质量作为最终译文质量的一部分，其质量好坏直接影响最终翻译质量。因此我们倾向于选择这样的模板——利用该模板进行翻译时，需翻译的各个片段可获得相对较高翻译质量。本文将短语翻译按质量分为三个层次，第一个层次包括英文字符、数字等不需要翻译的片段，以及在短语翻译记忆库中可以直接找到的短语，这种短语的译文可信度较高；第二个层次是能够在短语模板库或句子模板库中找到匹配模板的短语，这种短语由于结构化较强所以也具有较好的翻译质量，第三个层次是通过逐词翻译获得的译文，这样的短语具有的可信度一般比较低。本文定义短语翻译质量评价公式为：

$$PhraseScore(F(S',SP)) = \left(\sum_{1}^{phraNum} Phrase\_Translate\_Score \right) / Phrase\_Num \tag{6}$$

待翻译句子长度越长，被模板完全覆盖的概率越小，待翻译句子长度越短，误选模板的概率越大（一个句子可能有多个候选模板）[8]。因此本文将候选模板覆盖质量作为模板选优的一个因素。定义模板覆盖质量评价公式为：

$$MatchScore(F(S',SP)) = MatchNum \times (MatchNum/|SP|) \times \theta_1 \times \theta_2 \tag{7}$$

其中当S'中匹配的固定部分占模板SP中固定项的2/3以上时$\theta_1 = 1$，否则$\theta_1 = 0$，当S'中可变部分不存在删除操作时$\theta_2 = 1$，否则$\theta_2 = 0$。

对于每个F(S',SP)∈FSet(F',SP)，最终确定该映射关系的得分公式为：

$$Score(F(S',SP)) = \alpha \times DisScore + \beta \times PhraseScore + \gamma \times MatchScore \tag{8}$$

选择FSet(F',SP)中得分最高的F(S',SP)，若有多个得分相同则选择分数最高中的任一个，生成译文。

2.5 译文生成

利用F(S',SP)和源语模板SP与目标语模板TP之间的对齐关系，生成变换后句子S'和目标语模板TP之间的映射关系F(S',TP)。再根据F(S',TP)、短语翻译引擎生成最终的译文。译文生成规则如下：

⑴S'中未映射的部分f_i，首先利用短语翻译引擎翻译f_i，然后把f_i的译文插入到目标译文模板TP中,本方法中插入到f_i前方第一个映射不为空的片段后。

⑵S'中映射到模板TP固定项的片段f_i，不做任何处理；

⑶S'中映射到模板TP可变项的片段f_i，首先利用短语翻译引擎翻译f_i，然后用的f_i译文替换TP中的对应模板可变项；

(4)目标模板 TP 中未被映射的可变模板项 T_j，把 T_j 从 TP 中删除；如果 T_j 的右邻为固定部分且未被映射，则需要同时把 T_j 的右邻的固定部分也从 TP 中删除。

3 实验结果及分析

3.1 实验数据

实验使用的双语模板库包含 3000 对具有对齐关系的模板，数据为机器制造领域。我们分别构造了 100 句中文和 100 句日文进行测试，每个测试句子只提供一个标准答案。中文句子平均长度为 18.2 个字符，日文平均长度为 31.6 个字符。

3.2 评价方法

本文将 BLEU 值和系统翻译时间作为评价指标。BLEU 是一种基于 N-Gram 的自动评测方法，它将译文跟参考译文进行 N-Gram 的比较得出译文质量的评价分数。

3.3 实验结果及分析

在本文实验中，将基于浅层句法技术抽取句子模板，进行模板-模板匹配的系统作为 Baseline 系统，并构建基于文本-模板直接匹配的翻译系统 PBMT_A（无模板选优模块）和翻译系统 PBMT_B(有模板选优模块)。硬件资源：(CPU：2.2G，内存：2G)。表 1 给出了 Baseline 和 PBMT_A 在中到日(C2J)和日到中 (J2C)翻译结果上的 BLEU 值和系统时间开销：

表 1 Baseline 和 PBMT_A 结果比较

翻译方向	C2J		J2C	
	BLEU	翻译时间(s/句)	BLEU	翻译时间(s/句)
Baseline	0.3409	0.04	0.2662	0.09
PBMT_A	0.3914	0.04	0.3399	0.10

从上表中我们看以看出，基于文-模板直接匹配方法的系统在中日和日中翻译上分别提高了 5.05%和 7.37%。Baseline 系统在日到中翻译方向性能低于中到日，说明基于文本-模版直接匹配的 PBMT 在受限领域内取得了较好的性能。通过分析实验结果发现，浅层句法分析器在模板抽取上效果并不理想,因此导致 Baseline 系统性能较低。

表 2 分别给出了 PBMT_A 和 PBMT_B 在中日和日中翻译结果上的 BLEU 值和系统时间开销：

表 2 PBMT_A 和 PBMT_B 结果比较

翻译方向	C2J		J2C	
	BLEU	翻译时间(s/句)	BLEU	翻译时间(s/句)
PBMT_A	0.3914	0.04	0.3399	0.10
PBMT_B	0.4381	0.06	0.4016	0.20

从上表中可以看出，系统 PBMT_B 较 PBMT_A 在性能上有了很大提高。其中在中到日翻译方向上结果提高 4.67%，在日到中翻译方向上结果提高 6.62%。通过统计候选模板数随编辑距离分布变化如表3，模版选优模块的加入是 PBMT_B 较 PBMT_A 性能提高的主要原因。而在时间开销上，添加模版选优模块的 PBMT_B 系统在中日和日中方向上分别达到了 0.06s/句和 0.20s/

句，该系统在时间性能上仍需要进行改善。

表3 候选模板数随编辑距离变化分布表

编辑距离	0	≤1	≤2	≤3
中到日	1.48	6.72	16.54	27.63
日到中	2.79	4.16	15.18	23.94

从实验结果可以看出中到日的翻译结果在 BLEU 值上高于日到中的翻译结果，在翻译效率上中到日也明显好于日到中。分析原因如下：第一，日语可以借助词形变化和助词作用表达词语在句子中的意思，所以次序相当自由，因此对汉语生成时的语序调整造成了很大困难；第二，测试数据中，日文句子的平均长度较中文长且日文模板较为复杂，在通过路径方向矩阵解析映射关系时耗费的时间较多。

通过分析翻译结果发现，短语翻译质量对句子最终翻译结果有较大的影响。在短语翻译中，受资源所限，日文片假名的翻译质量较差。

实验表明，本文提出的文本-模板直接匹配算法将待翻译句子和实例库中的模板直接进行匹配，利用这种算法建立的 PBMT 系统，在受限领域翻译中取得了令人满意的性能。

4 结论

本文提出了一种面向机器翻译的文本-模板直接匹配算法。该算法可绕过模板抽取步骤，将待翻译句子和实例库中的模板直接进行匹配，以避免复杂的语法分析。利用这种算法建立的模板机器翻译模型，首先将待翻译句子根据源语模板进行变换，计算变换后句子和源语模板之间的相似度获得映射关系，并利用源语模板和目标语模板之间的对齐关系获得变换后句子和目标语模板间的映射关系，最终生成翻译结果。实验表明，利用这种算法建立的模板机器翻译引擎的翻译质量取得了令人满意的效果。为了进一步提高翻译质量，在以后的研究中，我们将进一步研究长句的多模板匹配，译文生成中由于插入和删除造成的调序及短语翻译的准确度等问题。

参考文献

[1] Nagao M.A. A framework of a mechanical translation between Japanese and English by analogy principle. Artificial and Human Intelligence, 1984,173-180.

[2] Hiroyuki KAJI, Yuuko KIDA, Yasutsugu MORIMOTO. Learning Translation Templates from Bilingual Text. Proceedings of the 14th conference on Computational linguistics, 1992, 672-678.

[3] Halil Altay Güvenir, Ilyas Cicekli. Learning Translation Templates From Bilingual Translation Examples. Infromation System, 1998,23(6),353-363.

[4] Brown R.D. Automated Generalization of Translation Examples. Proceedings of the 18th International Conference on Computational Linguistics, 2000, 125-131.

[5] 张学,黄德根.EBMT 翻译系统中模板的抽取与匹配.全国第八届计算语言学联合学术会议, 2005,681-683.

[6] 张俐. 面向奥运新闻的汉日机器翻译系统研究与实现.东北大学博士学位论文,2006.

[7] Levenshtein. V. Binary codes capable of correcting deletions, insertions and reversals. Doklady Akademii Nauk SSSR, 1965, 163(4):845–848.

[8] Nirenburg, Domashnev S. C, Grannes D. J. Two Approaches to Matching in Example-Based Machine Translation. TMI, 1993, 47–57.