人工书写翻译模板在统计机器翻译系统中的应用*

蔡舒 吕雅娟 刘群

中国科学院计算技术研究所，中国科学院智能信息处理重点实验室，北京，100190

{caishu,lvyajuan,liuqun}@ict.ac.cn

摘要： 人工书写的翻译模板包含语言学知识，能够准确捕捉语言对之间的对应关系。尽管这些模板的准确度高，但它们没有概率信息，难以被应用到统计机器翻译系统中，而且匹配时容易发生冲突。本文提出了一种将人工书写翻译模板应用到统计机器翻译系统中的方法，在统计机器翻译解码的过程中匹配模板，利用对数线性模型选择翻译。这种方法不仅提供了统计机器翻译系统高质量的翻译模板，还利用统计机器翻译系统的框架解决了模板匹配冲突问题。实验表明，这种应用提高了现有统计机器翻译系统的翻译质量。

关键词： 统计机器翻译 翻译模板 人工书写模板

Application of Linguistic Translation Templates in Statistical Machine Translation Systems

Shu Cai, Yajuan Lü and Qun Liu

Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

E-mail: {caishu, lvyajuan, liuqun}@ict.ac.cn

Abstract: Linguistic translation templates, which are usually written by humans, could capture the correspondence relationship between language pairs accurately. Although these templates have high accuracy, it is not easy to incorporate them in statistical machine translation systems since they do not have probabilities, and there are often conflictions when matching them. This paper presents a method to apply the linguistic translation templates to existing statistical machine translation systems, by matching linguistic translation templates during the decoding process, and choose the translation hypotheses based on log-linear model. This method not only provides the system with high-quality translation templates, but also takes advantage of the statistical machine translation systems to solve the confliction problem. Experiments show that this application could improve the translation quality of existing statistical machine translation systems.

Keywords: statistical machine translation, translation templates, linguistic translation templates

1 引言

随着国内外政治、经济、文化、科技等方面交流的日益频繁，各个国家之间的交流越来越重要，对语言翻译的需求越来越多。因此，近年来机器翻译在国内外得到了蓬勃发展，机器翻译系统也逐渐走向实用化。

能够进行机器翻译的前提之一是能从语言现象中总结出指导翻译过程的知识，这些知识描述不同语言之间的对应关系。在机器翻译的过程中，计算机“学习”到这些知识和它们的

*本文承自然科学基金项目(项目号 60736014, 60873167)和国家 863 重点项目(项目号 No. 2006AA010108)的资助。

应用范围，并利用这些知识进行翻译。

翻译模板是这些知识的一种合理表示方法，它通常的表示形式为分别用两种语言书写的包含常量和变量的字符串，这些字符串的应用限制，以及它们之间的对应关系。

从技术层面上说，机器翻译系统可以分为基于统计的机器翻译系统和基于规则的机器翻译系统两大类，它们获取并利用的知识都可视为翻译模板。前者利用统计方法获取这些模板，通过计算概率决定实际使用的模板；后者使用人工（通常是语言学家）编写的模板，通过各种规则的限制决定实际使用的模板。

基于统计的机器翻译系统通常利用双语平行语料作为翻译知识源，利用词对齐的限制，从中学习指导翻译的模板，使用带概率的模型来刻画翻译过程。它的优点是获取翻译模板的过程是自动的，不需要人工干预，因此可以在短时间内构建，易于转换语言对和改变领域。统计机器翻译系统中的模板冲突很容易解决，因为翻译模型给模板赋予概率，计算分数。但是基于统计的机器翻译系统也存在缺点。首先，词对齐本身就存在很多错误，获取到的模板通常包含很多噪声，翻译出的句子有时令人难以理解；其次，由于词对齐的限制和模板数量的限制 [Och et.al, 2004, Chiang 2007]，抽取的模板通常只能描述句子的一部分，缺乏概括性。所以，在翻译较长句子的过程中，如果源语言和目标语言的词序不一致，缺乏精确概括的模板指导，统计机器翻译系统很容易出现错误。

尽管在各种机器翻译系统的评测中，统计机器翻译系统取得了优秀的成绩，基于规则的机器翻译系统仍然在实用机器翻译系统中占有重要的位置。基于规则的机器翻译系统的翻译模板通常是人工书写的，包含语言学家总结出的知识，它的优点是规则噪声少，匹配精度高，具有概括性，翻译结果容易理解，规则数量较少。在实用机器翻译系统中，用户对翻译的准确度要求通常较高，这样的翻译模板刚好满足他们的需求。缺点是构建一个完全基于规则的机器翻译系统需要大量的时间和人力物力，并且不容易转换语言对和领域。当模板匹配发生冲突时，需要添加更多的规则和限制来解决。

可以看出，这两种机器翻译系统的优点是互补的。基于规则的机器翻译系统的翻译模板精确性和概括性较好，但匹配中容易发生冲突，构建成本较高；基于统计的机器翻译系统能够轻易解决模板冲突问题，构建成本低，但是不容易获取高质量的模板。如下的模板就很难用统计方法获取到：

模板源语言部分：一种具有 X1 作用的 X2 及其制备方法

模板目标语言部分：A X2 having X1 effects, and its preparing method

这个模板能够覆盖待翻译句子的主干内容，描述复杂的双语之间的顺序关系，并且能够泛化到多种结构相似句子的翻译。例如“一种具有补血作用的药物及其制备方法”，“一种具有防止脱发作用的冲剂及其制备方法”等。模板还可以带有附加匹配条件，提高模板匹配的准确率。例如限定 X1 必须是以名词结尾的部分。人工书写的模板较容易添加这些限制，而这些限制在通过统计方法获取的模板中却很难被归纳出来。

在面向特定领域的机器翻译中，人工书写的规则通常比较容易归纳，这些融合了语言学和特定领域知识的模板具有极高的价值。机器翻译系统的用户也往往希望自己能添加翻译模板，使翻译系统的性能得到提升。但是使用统计机器翻译系统时，这些知识往往不能被有效利用，因为这些规则的概率难以估计，且容易与用统计方法获取的知识冲突。

本文提出了一种将人工书写的模板集成到实用统计机器翻译系统中的方法，包含模板匹配算法以及改进的统计机器翻译解码方法，既有效利用了人工书写模板所包含的知识，又利用统计机器翻译系统的特点解决了模板冲突问题。

本文按如下方式组织：第 2 部分介绍相关工作，第 3 部分描述使用的模板匹配算法，第 4 部分介绍加入人工模板的统计机器翻译解码算法，第 5 部分是实验结果和分析，第 6 部分对研究做了小结，并描述了下一步的工作。

2 相关工作

最初的基于自动抽取模板的翻译研究是在基于实例的翻译研究基础上进行的。[Takeda et.al 1998] 给出了翻译模板的一种形式化定义，并给出了相应的翻译算法和算法复杂性的理论证明。这种翻译模板依赖词典等决定模板两部分的对齐关系。

统计机器翻译中的词语对齐出现以后，为统计方法自动抽取模板提供了基础。[Och et.al, 2004, Chiang 2007, Liu et.al 2006] 描述了目前一些主流统计机器翻译系统中翻译模板的抽取方法，其中[Liu et.al 2006]还结合了自动句法分析的结果。这些模板数量巨大，匹配费时，在实际应用中较难使用。通常的做法是在抽取模板的双语短语长度等方面做启发式限制，而这样的限制是以牺牲翻译质量为代价的。

[付雷等，2007] 尝试解决融合人工书写句型模板和统计机器翻译融合的问题，首先利用正则表达式匹配人工书写的模板，选择模板匹配结果，然后将匹配后的句子送入统计机器翻译系统顺序翻译。这种方法为现有统计机器翻译系统提供了一种融合语言学知识的方法，但是融合后的统计机器翻译系统只能顺序解码，限制了统计机器翻译的质量，而且匹配冲突时没有结合上下文选择正确的模板，而是根据长度优先匹配原则选择。

本文提出的方法针对特定领域的实用统计机器翻译系统，在统计机器翻译系统使用的翻译模板中加入少量人工书写的高质量模板，利用了这些模板包含的语言学知识，使用统计机器翻译框架解决模板匹配冲突问题。不同于[付雷等，2007]，我们在统计机器翻译解码的过程中匹配模板，保留了所有模板可能的匹配，在翻译过程中根据其他特征信息选择模板翻译，将人工模板的选择应用过程集成到统计机器翻译的对数线性模型中，对于不应用模板的部分，统计机器翻译系统可以根据统计翻译模型正常调序。

3 模板匹配算法

3.1 人工书写模板简介

在本实验中使用的人工书写模板详细定义如下：

由待翻译的源语言和目标语言组成的结构，每部分又分为“模板的常量”部分和“模板的变量”部分，其中的“模板的常量”部分为每一部分中去掉“模板的变量”部分后剩下的部分，而“模板的变量”部分的完整定义形式及含义如下：

源语言中模板变量部分的完整形式为：　　##N [m ,n] {+/－word}

目标语言中模板变量部分的完整形式为：　　##N

各部分的含义如下：

<1>　**##N：**　　句子中可以被泛化成变量部分的标志符，N 从 1 开始编号，源语言中的##N 与目标语言中的##N 一一对应

<2>　**[m, n]：**　　变量部分的长度限制。表示源语言中被泛化的变量部分的长度必须在某个范围之内。有以下几种变种形式

[m, n]　表示 m=<变量的长度<=n

[m,]　表示 m=<变量的长度

[,n]　表示 0=<变量的长度<=n

[0]　表示对变量的长度没有限制

<3> **{+/－word}：** 变量部分是否必须含有或不含有某些词语的限制。表示源语言中被泛化的变量部分必须含有或必须不能含有某些词语。词与词之间用空格分割，且“+”与“－”不能同时出现。有以下几种变种形式

{+word} 表示被泛化的变量部分必须同时含有 word 中的每一个词

{-word} 表示被泛化的变量部分必须不含有 word 中的任何一个词

{0} 表示对被泛化的变量部分对是否含有哪些词没有限制

一个汉英句型模板的实例如下：

- ##1[0]{0}和其制备方法及在##2[0]{0}中的应用
- ##1 for ##2 and its preparation and application

其中第一句称为源语言部分，第二句称为目标语言部分。模板可能还包含其他信息，如创建人，创建时间，模板序号等。

能够匹配该句型模板的句子实例如：

一种药物和其制备方法及在临床中的应用

3.2 模板匹配算法实现

给定一个人工书写模板的集合，在统计机器翻译过程中，对每一个待翻译片段，我们首先尝试匹配人工书写的模板。

在我们的系统实现中，采用对模板常量项建立索引，根据索引匹配片段，然后检查变量限制的方法来进行模板匹配。对于一个待翻译片段匹配多个模板和一个模板在同一翻译片段中有多个匹配位置的情况，这种算法都能返回完整的结果。

注意到人工书写的模板是没有分词信息的，而统计机器翻译系统中的输入句子是分好词的，匹配结果可能和模板边界有冲突。例如，下面的句子片段：

- 这 一 种子 和 它的 功能

如果不考虑分词，除了“这”以外的部分可以匹配下面的模板源语言端：

- 一种##1[0]{0}和##2[0]{0}

但是，考虑到分词的话，“种子”是一个词，它或者属于常量部分，或者属于变量部分，不应该同时属于常量和变量。所以这个句子片段不能匹配该模板。

对此，我们采用尊重分词的匹配方法，对于未分词前能匹配句型模板，但分词结果和模板边界有冲突的情况，认为与该句型模板不匹配。

匹配的结果可以表示为由模板序号，模板各常量的匹配位置，首尾变量的可能位置等组成的结构，解码器根据这个结构，在解码过程中选择需要使用的模板。

下图简要描述了模板匹配的过程。

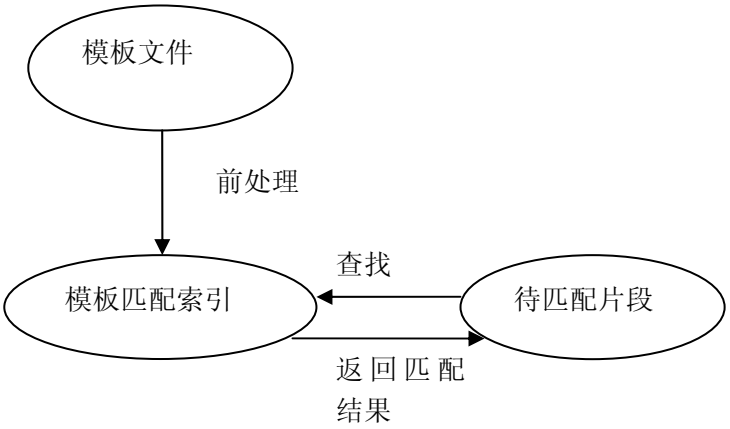

图 1 模板匹配过程图

首先对模板文件进行前处理，分离出源语言部分和目标语言部分，以及模板序号等一些附加信息。目标语言部分留待解码时使用，源语言部分分离出常量和变量部分，对各常量部分建立索引，写入索引文件。

在实际使用中，模板文件只需要读取一次，而建立的索引文件可以反复使用。索引以各常量部分为索引键，值为含有该常量的模板序号。

经过分词的句子有 n 个词，按 0,1,..n-1 编号，待匹配片段可以用一个词序号的二元组表示，两个词序号分别代表片段的开始词和结束词。

若一个句子能够匹配某序号模板的所有常量部分，则检查其变量部分，如果符合模板限制，则认为句子匹配此模板。需要注意的是如果变量位于模板的开头或结尾，如模板”##1[0]{0}和##2[0]{0}”，需要根据变量的要求确定匹配的长度，具体实现中对首尾变量的匹配返回一个词序号的范围。变量部分不能匹配为空。

最后返回解码器的是该句子匹配的模板序号以及各常量和变量的匹配位置，以供解码过程使用。

4 融合人工书写翻译模板的统计机器翻译解码

4.1 对数线性模型

主流统计机器翻译系统采用对数线性模型[Och et.al 2002, Koehn et.al 2003]作为解码器基本框架。对源语言 f 翻译成 e 的情况，翻译概率 P(e|f)可以如下表示：

$$P(e \mid f) = \frac{\exp[\sum_{1}^{M} \lambda_m h_m(e, f)]}{\sum_{e'} \exp[\sum_{1}^{M} \lambda_m h_m(e', f)]}$$

其中 M 是翻译系统的特征数，$h_m(e,f)$是第 m 个特征函数，λ_m 是它的权重。系统选择使 P(e|f)最大的 e 作为最佳翻译。

对于每个翻译模板（短语表可视为只含常量的翻译模板），有 4 个特征：p(e|f), p(f|e), lexp(e|f),lexp(f|e)，分别代表给定 f，翻译成 e 的概率；给定 e，翻译成 f 的概率，以及这两个方向的词汇化概率.[Koehn et.al 2003]

人工书写翻译模板也是一种翻译模板。要在统计机器翻译系统中使用人工书写翻译模板，可以将它视为一个具有较高概率的翻译模板。

4.2 融合人工翻译模板的解码方法

下图是融合了人工书写翻译模板的统计机器翻译系统的总体流程图。

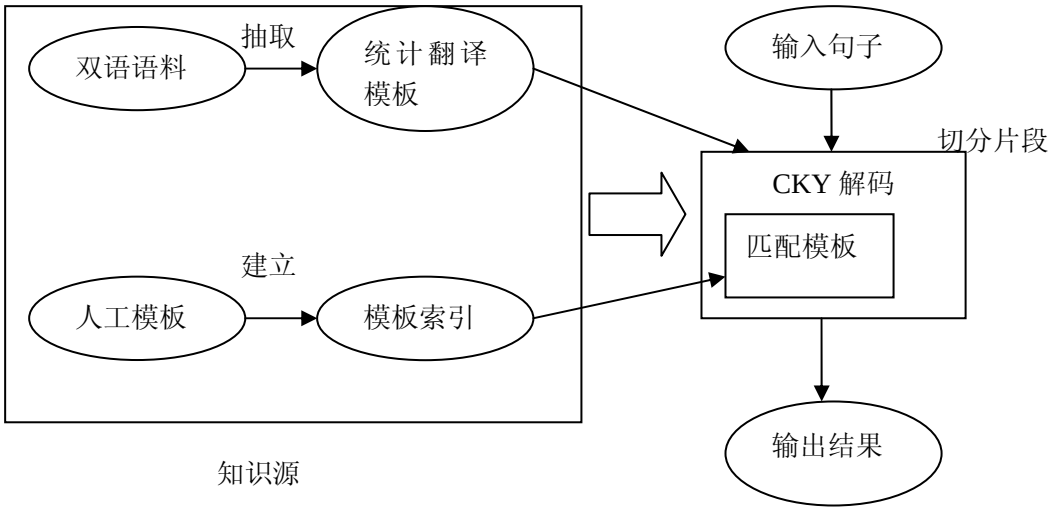

图 2 融合人工书写翻译模板的统计机器翻译总体流程图

我们将人工书写的翻译模板视为统计机器翻译系统的一个附加知识源，匹配模板的过程是在统计机器翻译系统 CKY 解码过程中添加的一个部分。

CKY 解码算法是自下而上查找翻译模板的过程，利用动态规划算法依次生成翻译假设，搜索的跨度从小到大，这个跨度也即模板匹配中的待匹配片段的跨度。对每个大于模板索引中最短模板长度的跨度，我们尝试匹配模板，以及确定模板中变量在待翻译跨度中的位置。

由于模板的跨度一定比模板中包含变量的跨度大，所以当解码到模板的跨度时，模板包含变量的候选翻译一定已经选出，用最高分数的变量的翻译替换模板中变量的翻译，就可以得到模板的翻译。如果用多个可能变量翻译替换模板中变量的翻译，就可以得到多个模板的可能翻译。

对于多个模板匹配或者多个位置匹配同一模板的问题，虽然翻译模板特征的概率都是 1，但是它们的其他特征，特别是语言模型特征不一样，造成最终生成的翻译假设的分数差别。由于根据特征打分考虑了多方面知识的因素，所以利用统计机器翻译的概率计算解决了模板冲突的问题。如果模板确定性不高，可以根据经验值将模板赋予的概率调低，这样选出的最好翻译有更大可能性是不使用模板的翻译。

本算法采用提供给解码器所有匹配的模板作选择，通过计算特征函数的加权分数来选择候选翻译的方式来进行解码，由于解码器有多个代表不同知识源的特征，用这些特征加权得到的分数给使用模板的翻译假设排序，而不是单纯地依据匹配长度或者其他启发式标准，使得模板的冲突选择更加精确，

这种将匹配模板融合于统计机器翻译解码过程的方式扩大了统计机器翻译候选的选择，并且给更有可能是候选翻译的人工书写模板赋予了更高的优先级，使统计机器翻译系统能够

将从统计翻译模板中学习到的知识与人工书写模板知识有机地结合，提高了翻译质量。

加入人工模板匹配和解码的 CKY 算法伪代码如下：

设句子长度为 n，每次搜索的跨度为(i, j)， H(i, j)代表所有(i,j)跨度的可能翻译。Rule 表示统计翻译模板（短语表等）。对于每一个 j>i 的跨度，进行解码前先匹配模板，生成候选翻译，最后输出(0,n-1)跨度的分数最高的翻译假设。

```
For i=1 to n
    If rule: A-> Wi exists
        Add Wi to P(i,i) // 因为没有模板只包含一个常量，对于所有的只含一个词的跨度，不匹配模板
 For i=1 to n
    For j=i-1 to 0
        If(j,i) matches some template
            Replace the variable translation in the template and add a new translation to hypotheses//加入一个新的使用模板的候选翻译
        Else
            Decode as usual
Output: the translation with the highest score(Or nbest translations with the highest scores)//输出按分数排序的翻译结果
```

图 3 融合模板匹配的改进 CKY 算法伪代码

5 实验与分析

我们的实验是面向专利翻译领域的，这个领域的句子具有一定的规律性，能够用人工书写模板描述。系统采用的是传统中药领域的专利翻译语料，训练集为 12 万句句对的双语平行语料，从训练语料中提取了开发集 300 句，测试集 200 句。开发集和测试集中每个句子的参考译文都为 1 个，语言模型采用 4 元语言模型，用训练集的英文部分训练。语料的中文部分用 ICTCLAS [Zhang et.al 2003] 分词。

我们使用基于最大熵括号转录语法的解码器[He et.al 2006, Xiong et.al 2006, Xiong et.al 2008]作为实验用解码器。该解码器在基于短语的统计机器翻译解码器基础上加入了使用最大熵分类器训练的重排序模型，具有较强的调序功能，也具有一般统计机器翻译系统的特性。

该解码器使用基于 CKY 算法的解码方法，使用以下特征：语言模型特征，单词数惩罚特征，短语数惩罚特征，翻译模板特征，重排序特征。在模板匹配实验中，加入人工书写的模板后，所有的人工模板的翻译模板特征被设为 1，并且使用了人工模板时，短语数特征改变为模板数+短语数特征。这体现了模板是一个“可信度较高的短语”的思想。

我们使用最小错误率训练[Och et.al 2003] 训练统计机器翻译模型的参数，用机器翻译中常用的评价标准 BLEU [Papineni et.al 2001] （大小写敏感）来评价翻译结果。

实验中使用的模板为人工从训练集中归纳总结的模板，共 104 句。表 1 显示了开发集和测试集模板的匹配情况。

	无任何模板匹配	仅一个模板匹配	有多种匹配方法	总共
开发集	147	18	135	300
测试集	99	12	89	200

表 1 开发集和测试集模板的匹配情况

可以看出，开发集和测试集都约有一半的句子没有模板能够匹配，在能够匹配上的模板中，大多数都有多种匹配方法，导致模板冲突问题。

我们使用不加模板的实验作为 baseline。我们重复了[付雷等，2007]的实验（实验一）以对比结果，先用正则表达式方法匹配模板，再将匹配好的句子送入解码器，解码器不调序。

实验二在解码过程中加入匹配模板的过程，用统计机器翻译方法决定要使用的模板和翻译结果。

实验结果如下表所示：

	开发集	测试集
baseline	27.29	25.92
实验一	28.22	26.73
实验二	**29.31**	**27.76**

表 2 实验结果

从实验结果可以看出，使用人工书写的模板后（实验一和实验二），开发集和测试集的 BLEU 值都有提高。实验一开发集比 baseline 提高了 0.93 个点，测试集提高了 0.81 个点。实验二用本文中的方法将人工书写模板应用于统计机器翻译系统，开发集在实验一的基础上提高了 1.09 个点，测试集提高了 1.03 个点。

从表 1 的统计数据可以看出，开发集和测试集的句子结构组成对模板匹配基本是一致的，能够匹配的句子占约 50%，大部分匹配了多个模板或者位置。这从一个侧面说明融合人工书写模板的解码方法对于解决模板匹配冲突的重要性。同时，加入模板匹配和解码后，统计机器翻译系统利用了可信度较高的知识，翻译质量得到了提高。

6 总结与展望

本文提出了一种将人工书写的翻译模板应用于统计机器翻译系统中的方法，包含模板匹配和选择模板两部分。这种方法在统计机器翻译系统中加入了较为可信的人工书写翻译模板指导翻译，同时利用了统计机器翻译系统的框架，结合上下文特征，在翻译解码过程中动态匹配和选择要使用的模板。实验证明，这种方法融合了基于规则和基于统计的机器翻译系统的优点，在实用机器翻译系统中提高了翻译质量。

在本文中，所有的人工书写翻译模板的可信度被视为相同。在下一步的工作中，我们将研究对人工书写翻译模板做出评价的方法，并在翻译过程中加入模板的可信度作为特征，期望进一步融合统计机器翻译系统和人工书写翻译模板的优点。

参考文献

Zhongjun He, Yang Liu, Deyi Xiong, Hongxu Hou, and Qun Liu. ICT System Description for the 2006 TC-STAR Run#2 SLT Evaluation. TC-STAR Evaluation Workshop, Barcelona, Spain, June 19-21. 2006.

Deyi Xiong, Qun Liu, and Shouxun Lin. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In Proceedings of COLING-ACL 2006, Sydney, Australia

Deyi Xiong, Min Zhang, Ai Ti Aw, Haitao Mi, Qun Liu and Shouxun Lin. Refinements in BTG-based Statistical Machine Translation. In Proceedings of IJCNLP 2008

Koichi Takeda, Pattern-based Context-Free Grammars for Machine Translation, Proceeding of the 34th ACL, 1998,pp.141-151

Franz Josef Och, Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 295-302.

Philipp Koehn, Franz Josef Och, and Daniel Marcu . Statistical phrase-based translation. In Proceedings of HLTNAACL, 2003, pages 127 – 133.

Franz Joseph Och and Hermann Ney. The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics, Vol 30, No.4, 2004, pages 417-449

David Chiang, Hierarchical phrase-based translation, Computational Linguistics, Volume 33, Issue 2, 2007, pp. 201-228

Yang Liu, Qun Liu, and Shou Xun Lin, 2006, Tree-to-string Alignment Template for Statistical Machine Translation. COLING-ACL 2006.

Franz Josef Och. Minimum error rate training for statistical machine translation. Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics, 2003. pages 160-167

Hua-Ping Zhang, Qun Liu, Xue-Qi Cheng, Hao Zhang and Hong-Kui Yu, Chinese Lexical Analysis Using Hierarchical Hidden Markov Model, proceedings of 2nd SigHan Workshop, August 2003, pp. 63-70

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, IBM Research, RC22176 (W0109-022) September 17, 2001

付雷，吕雅娟，刘群，一种融合了句型模板和统计机器翻译技术的翻译方法，第九届计算语言学联合学术会议，2007