# 面向小领域的可信机器翻译技术研究*

李贤华，于 淼，吕雅娟

中国科学院 计算技术研究所 智能信息处理重点实验室，北京 100190

E-mail: {lixianhua, yumiao, lvyajuan}@ict.ac.cn

**摘 要：**当前统计机器翻译的模型不断复杂、语料规模不断增加，但翻译质量仍是机器翻译实用化的瓶颈。在一些语料少、句子短、句式工整的小领域，可综合使用记忆库、词典、模板、规则、语言模型等资源，将基于统计和基于规则的机器翻译技术结合起来，实现小领域的可信翻译。本文使用层次短语模型，设计并实现了一个菜谱翻译系统。实验表明，本文设计的框架可有效利用多种资源在小领域上实现高质量的机器翻译。

**关键词：**小领域；可信机器翻译；层次短语

# Reliable Machine Translation Technologies in Specific Domains

Li Xianhua, Yu Miao, Lü Yajuan

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

E-mail: {lixianhua, yumiao, lvyajuan}@ict.ac.cn

**Abstract:** This paper proposes reliable machine translation technologies in specific domains. In some machine translation domains where the corpus is small, the sentences are short and the sentence structure is neat, employing translation memory, dictionary, templates and rules, combining the advantages of statistical machine translation and rule-based machine translation can achieve high translation quality. This paper researches on such technologies and builds up a system for translating menu. Results show that the system can translate in specific domains with good quality.

**Keywords:** specific domains; reliable machine translation; hierarchical phrase

## 1 引言

机器翻译是使用计算机将一种自然语言翻译为另一种自然语言的技术，是自然语言处理的热点与难点之一。近年来统计机器翻译研究取得了较大进展，但其翻译质量与语料规模和领域有很大关系，对于训练语料有限的小领域，翻译质量问题仍是机器翻译实用化的瓶颈。

当前统计机器翻译使用的模型日益复杂。从最初的基于字的模型[1]，到现在应用非常成熟的短语模型[2]，再到2005年后成为研究热点的句法模型[3][4]，模型训练和搜索的复杂度不断上升。句法系统对时空的要求很高，当前实用的句法系统较少。面向小领域的语料有句子简短、句式工整的特点，使得在小领域上构建句法翻译系统成为可能。

当前统计机器翻译使用的语料规模不断增大。在机器翻译的热门领域，例如新闻、科技等领域，动辄使用上百万甚至上千万的双语句对进行训练。然而，在面向小领域的机器翻译中，我们能获取的语料资源十分有限，有的领域的双语资源只有几千句对。如何使用有限的语料，达到较高精度的翻译质量，是本文研究的问题。

小领域翻译的一般特点为语料规模小、句子简短、句式工整，可以实现统一的、标准化的翻译。面向小领域的可信机器翻译，主要是为了实现在小领域（例如人名、菜谱、地址、机构名、职称职衔等领域）的高精度翻译。

* 基金资助：国家自然科学基金资助项目（60873167，60736014）。

本文后续内容组织如下：第 2 节介绍了层次短语模型的规则形式及解码算法，第 3 节介绍了小领域的可信机器翻译资源，第 4 节对层次短语模型做了一些改进，第 5 节实现了一个菜谱翻译系统，进行了实验及分析；最后是总结与展望。

## 2 层次短语模型

层次短语模型（Hierarchical Phrase Based Model）是由 David Chiang 在 2005 年提出的一种统计机器翻译模型[3][4]。它采用基于上下文无关文法（SCFG）的形式化语法结构，很多研究者将其归为基于句法的统计翻译模型。该模型不需要复杂的句法分析，可自动推导双语的形式化语法，具有很好的规则泛化能力和翻译调序能力，是目前性能最好的机器翻译模型之一。

### 2.1 层次短语模型的规则形式

层次短语模型使用的规则形式如下：

$$X \rightarrow < \alpha, \gamma, \sim >$$

其中，$X$ 是非终结符，$\alpha$ 是源端字符串，$\gamma$ 是目标端字符串，$\sim$ 是 $\alpha$ 与 $\gamma$ 的对齐关系。

层次短语模型定义了两种规则，一种是普通规则，另一种是粘贴规则。普通规则主要实现翻译和调序的功能，而粘贴规则主要实现复写和顺序翻译的功能。

普通规则是从语料库中自动学习的规则，它具有如下的形式：

$$X \rightarrow < \alpha_1 X_k \alpha_2, \gamma_1 X_k \gamma_2 >$$

表 1 是普通规则的具体举例。每条规则后面的四个数字分别为源端和目标端的正反向短语翻译概率、正反向词汇化翻译概率。

表 1 普通规则举例

| |
|---|
| 鳗鱼 ||| Eel ||| 0 -2.26868 -2.3979 0 |
| 炖 #X1# ||| Stewed #X1# ||| 0 -1.36432 -0.563118 -0.553385 |
| #X1# 焗 #X2# ||| Baked #X2# with #X1# ||| -0.313658 -0.741937 0 -0.930194 |

原始的层次短语抽取程序为了限制抽取的规则的数量，对规则进行了一定的限制[3][4]。本文所用的规则抽取程序沿用[3][4]中限制。

粘贴规则在第 4 节中进行介绍。

### 2.2 层次短语模型的解码算法

层次短语模型采用自底向上的 CKY 解码算法，并用 cube-pruning 算法[6][7]进行剪枝。候选翻译通过对数线性模型进行打分，最终得到最佳译文。

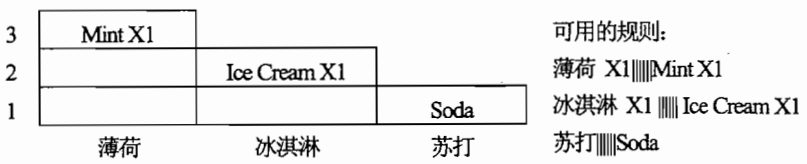


图 1 CKY 解码算法过程举例

CKY 算法的一个应用举例如下：假设输入的原文经过分词后，为“薄荷 冰淇淋 苏打”，能用的规则如图 1 中右半部分所示。画出如图 1 中左半部分的表格，自底向上、自左向右填写这个表格，填表的过程就是分析的过程。表格左部的数字，表示跨度的大小。最底层跨度为 1，有一条规则“苏打|||Soda”可用；第二层跨度为 2，有规则“冰淇淋 X1 ||| Ice Cream X1”，最上层有规则

“薄荷 X1|||||Mint X1”。通过规则的嵌套，可以将原文用这三条规则进行翻译，即将“薄荷 冰淇淋苏打”翻译为“Mint Ice Cream Soda”。

## 3 小领域可信翻译的主要资源

在小领域的可信机器翻译系统中，我们主要使用的资源有记忆库、词典、模板、规则和语言模型等。其中记忆库、词典和模板是经过人工整理的可信资源，而规则和语言模型则是从语料库中自动学习得到的。这些资源对小领域的翻译都有很重要的作用。

### 3.1 主要资源介绍

记忆库，是原文与译文一一对应的翻译资源，具体举例如表 2 所示。

表 2 记忆库举例

| |
|---|
| 白菜汤|||||Chinese Cabbage Soup |
| 红烧肉|||||Red-Cooked Pork |

词典，是词条与其候选翻译对应的翻译资源。词典支持一词多译，如果同一个词条有多个候选翻译，则多个候选翻译之间用“||”间隔。具体举例如表 3 所示。

表 3 词典举例

| |
|---|
| 乳酪|||||Cheese |
| 冰花|||||Rock Sugar||White Fungus |

模板是人工整理的可信度较高的翻译规则，有句首句尾约束。模板可以带一个或者两个变量，变量可以有含词的约束、含词的类别的约束、含词的词性约束等。具体举例如表 4 所示。

表 4 模板举例

| |
|---|
| $#X1# 炖 #X2#$ ||| Stewed #X2# in #X1# |
| $拔丝 #X1#$ ||| #X1# in Hot Toffee ||| #X1#: {word: 香蕉|| 苹果; class: 烹饪原料; pos: n} |

规则即第二节中介绍的使用规则抽取程序自动抽取的翻译规则。

语言模型我们使用 SRILM 的四元语言模型[5]。

### 3.2 可信资源的获取方法

由于小领域的语料规模较小，可信资源（记忆库，词典和模板）的质量对翻译质量有着重要的影响。其中记忆库主要来源于训练集的双语句对以及网络资源，获取方式比较简单。

词典和模板的主要来源为训练集语料等资源，需要人工校对，以保证其质量。

词典和模板的来源之一，是从训练语料自动抽取得到的规则表。规则表中抽取的规则有两种形式，第一种是不带变量的基本短语，第二种是带变量的层次短语。将这两类短语中质量较好的抽取出来，可分别整理为词典和模板。

规则质量的好坏，取决于其源端和目标端的正反向短语翻译概率 $p(e|f)$ 和 $p(f|e)$。如果规则的正反向短语翻译概率符合如下两个条件：

$$p(e|f)+p(f|e)\geqslant a \tag{1}$$

$$b\leqslant p(e|f)/p(f|e)\leqslant c \tag{2}$$

则该规则质量较好，可进一步人工确定是否整理为词典词条或模板。通过对规则表进行考察，我们设置 $a=1.5$，$b=0.8$，$c=1.2$。

词典和模板的其他来源，包括网络资源等，这些资源需要手工整理。

### 3.3 可信资源的使用方法

记忆库的使用比较简单。待翻译句子将首先查找记忆库，如在记忆库中存在该句子，则将其对应翻译取出，即得到译文，完成翻译。

原始的层次短语模型使用自动抽取的规则，通过CKY解码算法对输入句子进行翻译，不同的翻译之间通过特征的线性组合进行竞争。词典和模板是可信的翻译资源，本文将词典和模板一起加入规则表中，使其与规则一起参与竞争。由于词典和模板是可信资源，我们将其正反向短语翻译概率以及正反向词汇化翻译概率全部赋值为最高概率1。

词典的作用主要有两个：第一，将词典加入到规则表中，将其与规则一起自动参与解码器的竞争。词典词条的加入，减少了未登录词出现的可能性，可以提高翻译质量。第二，词典可以用来过滤规则。词典中的词条是具有高可信度的，在进行翻译时，如果某一个短语块包含词典中的词条的原文端，则该短语块对应的规则必须包含词典的目标端，否则该规则将被过滤掉。

模板则主要起到了增加可信规则的作用。将模板与模板中对变量的约束存储起来，并将模板加入到规则表中。当模板匹配到源端的片段的时候，将检查模板的泛化部分是否符合模板的约束条件。如果泛化部分符合模板的约束条件，则该模板将加入规则的竞争，否则，该模板在此片段不能适用，不能参加后继的竞争。

## 4 改进的粘贴规则

原始的基于层次短语的翻译模型使用了两条粘贴规则：

$$S \rightarrow < X_1, X_1 > \qquad \text{粘贴规则1}$$

$$S \rightarrow < S_1 X_2, S_1 X_2 > \qquad \text{粘贴规则2}$$

在实际翻译中，我们发现仅有这两条粘贴规则是不够的。由于层次短语使用了自底向上的CKY解码算法，如果底层短语没有对应的规则，会造成上层的短语也无法进行翻译。

如图2所示，仍然以“薄荷 冰淇淋 苏打”为例，可用的规则在图2右半部分中。


图2 CKY解码算法过程举例

按照常理，将三条可用规则顺序粘贴，即可得到原文的对应译文“Mint Ice Cream Soda”。然而，由于经典的层次短语模型中，源端的规则不允许有连续的变量出现，导致规则“冰淇淋||||Ice Cream”与“苏打||||Soda”无法顺序粘贴，填表过程无法继续，该词条无法翻译。类似这样依靠顺序粘贴就可得到译文的词条，也因为同样的原因无法翻译。

为了解决这个问题，我们在原有的模型中加入了第三条粘贴规则。这条规则允许源端的各个单元进行顺序粘贴，其形式如下：

$$S \rightarrow < X_1 X_2, X_1 X_2 > \qquad \text{粘贴规则3}$$

有了粘贴规则3之后，就可以进行顺序的粘贴了。

在经典模型中，出于加快解码速度的考虑，并没有使用该粘贴规则。由于现在的机器翻译的翻译领域一般为新闻、科技等领域，这些领域的句子较长，解码时间也相应较长。如果加入本文中的粘贴规则3，则会大大增大搜索空间，降低解码速度。而且，在这些领域上，由于语料丰富，能够抽到大量的规则，粘贴规则3的作用并不明显。然而，在面向小领域的翻译中，由于需要翻译的词条一般比较简短，增加该粘贴规则不会对解码速度造成很大的影响。而且，小领域的语料

一般较少，抽取的规则也有限，粘贴规则 3 的作用是比较明显的。

# 5 实验与分析

## 5.1 实验设置

本文搭建了一个面向小领域的可信机器翻译系统。系统各个特征的权重参数用最小错误率训练方法[8]在开发集上迭代得到，解码器候选翻译个数为 150，最终翻译的 N-best 值为 200。

## 5.2 实验数据

实验主要使用了两份菜谱语料，一份是北京市人民政府外事办公室提供的《中文菜单英文译法》，主要用来整理训练集、开发集和测试集，语料信息如表 5 所示；另一份是中文单语的《菜谱名单列表》，经整理后得到 29381 个菜名，主要用来整理词典、模板等资源。

表 5 语料信息统计

| | 句对数 | 平均句长（词） | 平均句长（字） |
|---|---|---|---|
| 训练集 | 1610 | 3.62 | 4.78 |
| 开发集 | 201 | 3.93 | 4.91 |
| 测试集 | 201 | 3.84 | 4.93 |

本翻译系统综合使用了记忆库、词典、模板、规则表和语言模型等资源。其中记忆库中并没有加入任何词条，词典词条总共 2566 条，模板 517 条。

## 5.3 实验结果和讨论

本文设计了多组实验，分别测试词典、模板和粘贴规则 3 对系统的影响。基准系统为标准的层次短语系统。使用国际机器翻译评测中通用的评价指标 BLEU [9]对系统的翻译质量进行评价，大小写不敏感。在测试集上的结果如表 6 所示。

表 6 单独添加词典、模板或粘贴规则 3 后的 BLEU 值

| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|
| baseline | 0.5684 | 0.4556 | 0.3713 | 0.296 |
| +词典 | **0.6586++** | **0.5366++** | **0.4352++** | **0.353++** |
| +模板 | **0.6058++** | **0.4988++** | **0.4211++** | **0.3548++** |
| +粘贴规则 3 | 0.5662- | 0.4424- | 0.3473-- | 0.2867- |

由此可见，单独使用词典或模板，对于系统的翻译性能都有很大的提高。用 BLEU-4 测试，使用词典可以提高 5.7 个点，而使用模板则提高了 5.88 个点。单独使用粘贴规则 3，会使系统的 BLEU 值略有下降。其中原因是粘贴规则 3 的存在，影响了嵌套短语的使用。

表 7 资源组合使用后的 BLEU 值

| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|
| baseline | 0.5684 | 0.4556 | 0.3713 | 0.296 |
| +词典+模板 | **0.6807++** | **0.5674++** | **0.4777++** | **0.4104++** |
| +词典+粘贴规则 3 | **0.6515++** | **0.5314++** | **0.4387++** | **0.3763++** |
| +模板+粘贴规则 3 | **0.5855++** | **0.4773++** | **0.3964++** | **0.3619++** |
| +所有 | **0.6697++** | **0.5551++** | **0.4674++** | **0.421++** |

模板和词典有部分内容交叉存在，而粘贴规则 3 的存在，会使规则的使用情况发生一些变化，因此，我们做了表 7 中的实验，测试组合使用词典、模板以及粘贴规则 3 的效果。

表 7 表明，词典和模板组合使用后，相对 baseline 提升的 BLEU 值，并不等于单独使用词典或模板提升的 BLEU 值之和，但比单独使用词典或模板的 BLEU 值都高。这说明词典和模板有一部分作用是重合的。总体上，它们对系统 BLEU 值都有很大影响。

另外，虽然单独使用粘贴规则 3 的时候，在测试集上 BLEU 值略有下降，但是当粘贴规则 3 与词典组合使用的时候，系统的 BLEU 值有明显的提高，这说明在词典资源丰富的情况下，在小领域的翻译中，粘贴规则 3 是可以起到一定的作用的。

系统在综合利用词典、模板和粘贴规则 3 的情况下，测试集上的 BLEU-4 值提高了 12.5 个点，达到了 0.421。为了测试词典、模板和粘贴规则 3 的作用，这些实验中均没有使用记忆库。如果使用了记忆库，相信系统的性能又会有大幅度的提升。

在实际应用中，可不断补充可信资源，系统的翻译质量会得到不断提升。

## 6 总结与展望

本文提出了一种面向小领域的可信机器翻译系统框架。本框架使用层次短语模型，并综合利用了多种资源，能高质量地在小领域进行翻译。对菜谱翻译感兴趣的朋友，可访问搭建的菜谱在线演示系统：http://nlp.ict.ac.cn/demo/cookBook，系统根据译文的来源是否为可信资源自动给译文进行打分。欢迎大家提出宝贵的意见和建议。

小领域由于其语料少、句子短、句型工整等特点，可以实现标准化的翻译。但是由于小领域的语料较少，实现高精度翻译需要大量的细致的工作。接下来的工作包括：使用本文提出来的框架，针对小领域的不同特点，在多个小领域实现高精度的机器翻译。

## 参 考 文 献


[1] Peter.F.Brown, Stephen A.Della Pietra, Vincent J.Della Pietra, Robert L.Mercer. The Mathematics of Statistical Machine Translation: Paramater Estimation[J]. Computational Linguistics, Vol 19, pages 263-311. 1993.

[2] Ye-Yi Wang and Alex Waibel. Modeling with Structures in Statistical Machine Translation[C]. In Proc. of COLING/ACL, pages 1357-1363. 1998.

[3] David Chiang. A Hierarchical Phrase-based Model for Statistical Machine Translation[C]. In Proc. of the 43th Annual Meeting on ACL, pages 263-270. 2005.

[4] David Chiang. Hierarchical Phrase Based Translation[J]. Computational Linguistics, Vol 33, pages 201-228. 2007.

[5] Andreas Stolcke. Srilm-an Extensible Language Modeling Toolkit[C]. In Proc. of the International Conference on Spoken Language Processing. Vol 2, pages 901-904. 2002.

[6] Liang Huang, David Chiang. Better k-best Parsing[C]. In Proceedings of 9th International Workshop on Parsing Technologies. Pages 53-64. 2005.

[7] Liang Huang, David Chiang. Forest Rescoring: Faster Decoding with Integrated Language Models[C]. In Proc. of the 45th Annual Meeting on ACL, pages 144-151. 2007.

[8] Franz Josef Och. Minimum Error Rate Training for Statistical Machine Translation[C]. In Proc. of the 41st Annual Meeting on ACL. Vol 1, pages 160-167. 2003.

[9] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation[C]. In Proc. of the 40th Annual Meeting on ACL, pages 311-318. 2002.