文章编号：1003-0077（2011）00-0000-00

翻译规则剪枝与基于半强制解码和变分贝叶斯推理的模型训练*

高恩婷[1,]，段湘煜[2]，巢佳媛[2]，张民[2]

（1.苏州科技学院电子与信息工程学院，江苏省苏州市 215011；2. 苏州大学计算机科学与技术学院，江苏省苏州市 215006）

摘要：统计机器翻译一般采用启发式方法训练翻译模型。但启发式方法的理论基础不够完善，因此，会导致翻译模型规模庞大以及模型参数精确率不高。针对以上两个问题，本文提出一种基于变分贝叶斯推理的模型训练方法，形成更精确的精简翻译模型。该方法首先通过强制解码对齐语料，然后利用变分贝叶斯 EM 算法获得模型参数。本文的实验语料为 NIST 汉英翻译任务数据，实验结果显示，基于句法（基于短语）的统计机器翻译中，超过 95%（76%）的规则被剪枝，且 BLEU 值显著提高。

关键词：机器翻译；规则剪枝；半强制解码；变分贝叶斯

中图分类号：TP391　　**文献标识码：**A

Translation Rule Pruning and Model Training with Semi-Forced Decoding and Variational Bayesian Inference

Enting Gao[1], Xiangyu Duan[2],Jiayuan Chao[2], Min Zhang[2]

(1. College of Electronics & Information Engineering，Suzhou University of Science and Technology，Suzhou 215011，China; 2. School of Computer Science & Technology, Soochow University, Suzhou215006, China)

Abstract: SMT usually learns translation model with heuristics. This makes model size very big and model parameters potentially less accurate due to the poor theoretical justification of heuristics. This paper presents a variational Bayesian inference-based training method to address the two issues, targeting to learn a compact trans-lation model with more accurate translation probabilities. It is achieved by translation model parameter estimation using variational Bayesian EM over alignments obtained by forced decoding. Experimental results on the Chinese-English NIST translation data shows that our proposed method is very effective, resulting in more than 95% (76%) rule pruned out with significant performance improvement in Bleu score for syntax-based SMT and phrase-based SMT.

Key words: Machine Translaation; Rule Pruning; Semi-Forced Decoding; Variational Bayesian

1 引言

目前，统计机器翻译（SMT）通常采用启发式方法训练翻译模型。首先，利用启发式方法（包括 intersection, grow, grow-diagonal, grow-diagonal-final and union）在训练数据上构建双向词对齐；然后在词对齐的基础上抽取翻译规则，抽取方法同样也是利用启发式方法，通过设置不同的剪枝阈值，如最大规则高度/宽度、每个跨度的最大规则数目、结点数量等，控制翻译规则的数量；最后对于抽取出的翻译规则集进行概率计算，其概率通常定义为规则的相对频率。启发式方法的主要优点在于它的简洁性，便于理解和实现。因此，启发式方法广泛用于各项前沿研究中。

但是，启发式方法的理论基础不够完善，训练过程独立于解码和建模过程。因此，启发式方法的翻译模型非最优，且学习到的大量翻译规则都是冗余规则。针对这两个问题，本文

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（61373095）

采用变分贝叶斯 EM 算法(Variational Bayesian EM, VBEM)来训练翻译模型(Antoniak 1974，Blei 和 Jordan 2005，Kurihara 等 2007，Johnson2008)。VBEM 算法是传统 EM 算法的扩展，且两者的总体框架相似（Dempster 等 1977)。VBEM 算法的 E 步对训练集进行强制对齐，M 步更新模型。VBEM 算法和 EM 算法主要区别于目标函数，体现在 M 步如何对模型进行更新。VBEM 算法的优点在于，引入先验知识克服 EM 算法过估计“大规则”而忽视“小规则”的过度拟合问题。与启发式方法相比，VBEM 算法能得到最终在翻译派生路径中实际使用的翻译规则，可确保训练过程更符合解码和建模过程。因此，VBEM 算法能兼顾小规则集，并且使翻译概率更精确。在 NIST 汉英翻译任务上的实验结果显示本文方法有效。

2 相关工作

本节分别从两个方面介绍相关工作：SMT 模型剪枝和模型训练。

SMT 模型剪枝的相关研究（Iglesias 等 2009，He 等 2009；Frantzi 和 Ananiadou 1996；Wang 等 2010；Eck 等 2007；Johnson 等 2007）都采用基于某些统计量（或启发式）的方法来实现模型剪枝，其所使用的剪枝策略与翻译模型训练和解码相对独立。与之相比，本文没有利用启发式方法，而是通过训练过程和强制对齐来决定所需规则。由于使用了稀疏先验分布，本文学得的翻译规则较小，具有较强的泛化能力，并且因为强制对齐的应用，翻译规则由实际派生路径上抽取而得，与实际解码过程在统计意义上一致。

SMT 模型训练的相关研究中，存在两类方法不使用启发式方法进行模型训练。一类是基于 EM 算法的方法（Marcu 和 Wong 2002；Denero 等 2006；Marcu 和 Wong 2006； May 和 Knight 2007；Wuebker 等 2010)。另一类是基于贝叶斯学习的方法（DeNero 等 2008；Blunsom 等 2009；Blunsom 等 2008a；Blunsom 和 Osborne 2008b；Cohn 和 Blunsom 2009)。这两类方法与启发式方法的主要区别是如何获取对齐结构。启发式方法第一步先获得词对齐信息，第二步再在词对齐的基础上抽取更高层次的结构对齐，这些结构对齐必须和词对齐相一致，但无法保证合法的派生路径（解码路径)，从而使相对频率的估计也不准确。这两步之间在统计意义上相互独立，对齐结构的获取缺乏理论依据，从而引起了放弃启发式方法的研究，主要包括上述提及的基于 EM 算法的方法和基于贝叶斯学习的方法，这两类方法直接学得结构对齐，不使用割裂的中间步骤。

基于 EM 算法的方法的首先提出者是 Marcu 和 Wong（2002)，将词对齐模型（IBM 模型）扩展到短语对齐模型。但由于 EM 算法存在过度拟合问题，导致往往较长的短语对被抽取出来，在极端情况下整个句对会被作为一个短语对抽取出来。这个问题随后被深入地进行分析（Denero 等 2006)，并先后有相应的方法加以解决，但都是在较小规模上进行实验，没有证据表明这些方法具备进行大规模应用的能力。因此，基于 EM 算法的方法不能有效解决基于启发式方法的模型训练存在的问题。

基于贝叶斯学习的方法可通过引入稀疏先验分布调整过度拟合问题，得到的后验分布往往比 EM 算法得到的分布更加稀疏，即只有较为常用的对齐结构的概率较大，出现稀疏的概率峰值，而不是 EM 算法得到的比较平均的概率值。其中代表性的方法有针对树到串对齐结构的方法（Cohn 和 Blunsom 2009）和针对短语对齐结构的方法（Blunsom 等 2009)，实验表明具有较强泛化能力的对齐结构可以通过贝叶斯学习的方法获得。由于关注一步得到结构对齐，贝叶斯学习方法具有较高的复杂度。Blunsom 等（2009）提出了局部 Gibbs 抽样方法，可以避免对整个平行句对进行计算的较高复杂度，但是局部 Gibbs 抽样方法有较慢的混合（mixing）速度，即样本抽样会陷落在一个局部最优点附近而不易产生新的抽样。为克服混合速度慢的问题，块化的 Gibbs 抽样被应用于树到串对齐结构抽样上（Cohn 和 Blunsom 2009)，使得整个平行句对同时被抽样，易于脱离局部最优点。

本文利用变分贝叶斯推理(Variational Bayesian Inference)训练翻译模型，不仅克服了在结构对齐学习中容易出现的过度拟合问题，而且通过引入平均场（Mean Field）降低了推理算

法的复杂度。本建模方法不依赖于启发式方法（如限制规则数量，或通过词法概率进行平滑）。相较于 EM 算法，变分贝叶斯推理可以获得较为稀疏的对齐结构；相较于基于贝叶斯学习的抽样方法，变分贝叶斯推理性能更好且容易实现，能解决对大规模语料和长句进行参数估计时存在的问题。

3 基于变分贝叶斯推理（Variational Bayesian Inference）和强制解码的翻译模型剪枝和训练

本文所用方法的模型训练框架概述如下：

1) 利用传统的基于启发式方法的模型作为初始 Bootstrapping 模型。
2) 用现有的模型对训练语料进行强制解码。
3) 利用步骤 2 中强制对齐后的训练语料，更新现有的模型。
4) 在开发集上对模型权重调参，返回到步骤 2 直到收敛至最优。

本文不使用额外的语言学资源，而是通过简化训练过程来获得相应规则，并从训练语料中获得规则的相应概率。

3.1 变分贝叶斯推理（Variational Bayesian Inference）

变分贝叶斯推理根据先验知识能解决传统 EM 算法的过度拟合问题。变分贝叶斯推理寻找近似后验概率的分布（用 KL 距离度量），便于计算后验概率（Liang 和 Klein 2007）。贝叶斯学习系统通常将 Dirichlet 分布作为先验知识。因为贝叶斯学习系统简易且有效，本文对 Dirichlet 先验知识使用平均场（mean-field）变分贝叶斯 EM 算法来最大化后验概率。平均场（mean-field）是一个近似完全分解的变分推理。平均场（mean-field）便于理解和实现，包括两个步骤（Johnson 2008）：

1) E 步：参照传统 EM 算法计算期望值；
2) M 步：分为以下三步：
   a. 加入 Dirichlet 超参数α到期望C_r。
   b. 将$\alpha + C_r$代入到$\exp(\Psi(.))$，得到$C_r' = \exp(\Psi(\alpha + C_r))$，其中，函数$\Psi(.) = \frac{\partial \Gamma(x)}{\partial x}$易于计算，相关代码见脚注[1]。
   c. 利用C_r'更新模型。

由上述步骤可知，平均场（mean-field）在使用 digamma 函数得到 Dirichlet 先验知识的基础上，重新计算期望值，动机来源于更有效的规则重用。基于 EM 算法的 SMT 模型训练、翻译规则（通常规模小）能提高最终翻译概率，但难以泛化未知数据。平均场（mean-field）通过 Dirichlet 先验知识对难以使用的规则进行惩罚，这是本文的核心思想，而另一个则是强制解码。

3.2 利用平均场（Mean-Field）进行 SMT 模型训练

本文的剪枝与训练过程是强制解码与平均场（mean-field）算法的融合。主要包含以下两个步骤：

1) 使用启发式方法抽取传统模型进行自训练；
2) E 步：

[1]http://web.science.mq.edu.au/~mjohnson/code/digamma.c

a. 利用现有的模型对训练语料进行强制解码。

b. 根据传统 EM 算法计算每个规则的期望值C_r。

3) M 步

a. 计算规则r的翻译概率：

$$p(r)=\frac{\exp\left(\Psi\left(\alpha+C_r\right)\right)}{\exp\left(\Psi\left(\alpha+\sum_{r'i=1}^{K}C_{r'i}\right)\right)} \quad (1)$$

其中，α是 Dirichlet 的超参数，参考本文中 3.1 节的 2.b 步骤的。

4) 在开发集上，对更新的模型权重调参，返回到步骤 2 直到收敛至最优。

公式（1）和传统 EM 算法的关键区别在于：在 EM 算法中$p(r)=\frac{C_r}{\sum_{r'i=1}^{K}C_{r'i}}$。通过引入 Dirichlet 先验分布，EM 算法中的概率估计被平滑，形式上相当于对 EM 算法进行了加一平滑。另外一个区别是：本文所用方法中的训练算法不被限制于 E 步中的词对齐。

3.3 强制和半强制解码的对比

强制解码主要分三个步骤：首先训练得到所有在标准翻译系统中使用的模型，接着使用 MERT（Koehn 等 2007）方法在开发集上进行模型参数调试以获得良好的 BLEU 得分，再接着使用这些模型和参数在训练集上进行解码，解码路径包含着结构对齐信息。在这些结构对齐的基础上，我们可以重新估计翻译规则的概率，而使其它模型保持不变。上述三个步骤重复迭代，直至前后两次迭代之间的解码路径不存在显著差异。强制解码的优势在于使得模型训练和解码过程一致，克服了启发式方法的模型训练与解码过程割裂开来的缺点，具有统计意义上的理论基础。

强制解码也存在实际应用的不足：往往较长的翻译规则被保存在最终的解码路径上。这是因为解码路径上的分解的翻译规则越少，解码路径的概率越高，从而使强制解码倾向于使用较少的翻译规则来完成解码，导致较长的翻译规则被最终保留。为克服这个不足，Wuebker 等（2010）使用 leaving-one-out 方法对各个结构对齐的概率进行平滑，以降低过高的较少使用的对齐结构（即较长的对齐结构）的概率，提高过低的泛化能力高的对齐结构的概率。本文使用另外一种方法来克服此种不足，通过 3.1 节所述的变分贝叶斯引入稀疏先验分布，寻找泛化能力高的翻译结构。

此外，强制解码要求部分假设必须与参考相兼容，最终翻译必须与参考相一致。但由于翻译规则的抽取未必能覆盖整个平行句对，导致某些平行句对不能产生有效的解码路径。与西方语言翻译到英文相比，汉英翻译的这个问题更加严重，部分句对不能成功强制解码。本文在汉英实验数据上的实验结果显示，使用 Moses（Koehn 等 2007)强制解码的句对中只有 72.2%能成功解码。同样，使用重新实现的基于森林的树到串句法系统强制解码，显示只有 31.4%句对能成功解码。

为使强制解码达到 100%的成功率，且确保翻译结果与参考尽量相似，本文采用半强制解码的方法。该方法引入一个新的特征度量在解码过程中的部分翻译结果与参考译文的相似性，这个特征可由 WER（错误率）、PER（位置独立的 WER）或 BLEU（Papineni 等 2002）表示，并在部分训练集上调整特征权重。

4 实验结果和讨论

4.1 实验设置

本文在两个 SMT 系统上进行评估，这两个系统分别是基于短语的统计机器翻译系统 Moses（Koehn 等 2007）和重新实现的基于森林的树到串系统（Mi 等 2008，Zhang 等 2009），并在两个解码器上实现强制/半强制解码功能。以下是在两个系统的实验设置，NIST 2002 测试集作为开发集，而 NIST 2003 和 NIST 2005 测试集作为测试集。GIZA++（Och 和 Ney 2003）和启发式方法“grow-diag-final-and”被用于生成汉英双语词对齐，并在两个系统中采用默认特征。本文利用改良后的 Koehn’s MERT 训练器（Koehn 等 2007）作为 MERT 训练器（Och 2003），使用 Zhang 的实现（Zhang 等 2004）进行显著性实验，并采用区分大小写的 BLEU-4(Papineni 等 2002)进行翻译质量评估。

对于基于句法的系统，训练数据来源于 LDC NIST-MT 的子集，包含 3 万个句对。本文利用 SRILM 工具（Stolcke 2002）和改良后的 Knese-Ney 平滑方法（Kneser 和 Ney 1995）在训练数据的目标端建立 3 元语言模型，并在中文 CTB5.0 上训练 Charniak’s 分析器（Charniak 2000），对分析器修改后使其输出封装后的森林。

对于基于短语的系统，训练数据是 24 万个汉英句对（21 万个 FBIS 和 3 万个 NIST-MT 数据）。本文利用 SRILM 工具和改良后的 Knese-Ney 平滑算法，在训练语料和英文 Gigaword 的新华社语料上，对目标端训练得到 4 元语言模型。

4.2 规则剪枝的实验结果

为公平比较，本文首先用传统模型的过滤技巧删除部分规则，例如，每个源短语或树保留 20 个目标翻译，删除非功能词没有被翻译的语法规则等。上述技巧被广泛用于当前系统（Koehn 等 2007，Liu 等 2006），且被证明不会降低翻译的精确性。

在基于句法的系统中，本文用到的规则剪枝如下：

1) 利用最优维特比算法搜集所有规则来生成一个小规则集。
2) 在小规则集上重新调参和测试

表 1 规则剪枝（基于句法的系统）

	BLEU		**模型规模**
	NIST03	**NIST05**	
初始	0.2394	0.2208	856M
剪枝后	**0.2486**	**0.2301**	**30M**

表 1 显示本文所用剪枝方法的有效性，可以看出，剪枝后的模型规模由 856M 减小到 30M，缩小 856/30=**28.5** 倍（表明减少(856-30)/856=**96.5%**的冗余规则），且翻译的精确性明显提高（$p<0.01$）。主要原因是保留重要的规则同时删除大量的不良规则（见表 2 和表 3，主要是局部词汇化规则）。这说明对于基于句法的机器翻译系统，局部词汇化规则没有其它规则重要。由于更大的搜索空间使解码器能够搜索到最优的结果，所以最终导致搜索错误得到了减少。

表 2 显示规则集中多数为局部词汇化规则，而本文能够将其规模降低约 40 倍。局部词汇化规则由于具有较细的颗粒度，容易引起过适应问题，从而导致翻译模型的泛化能力变弱。本文所用方法尝试保留具有高度泛化能力的翻译规则，因而倾向于使用具体的词汇信息越少越好。可以看到，通过本文中所提出的剪枝策略，低泛化能力的局部词汇化信息被过滤掉。另外，不同类型的规则的减少率是不同的。这表明本文的方法能够自动检测不同类型的有用规则，并改变最终剪枝模型的分布。

表 2 不同类型规则的减少率

	初始	剪枝后	减少率
完全词汇化	662477	88749	**7.5**
局部词汇化	3498535	88079	**40**
非词汇化	152151	4744	**32**

表 3 规则类型分布

	初始	剪枝后	F1K	N03	N05
完全词汇化	15.4%	48.9%	48.5%	47.3%	47.9%
局部词汇化	81.1%	48.5%	45.6%	45.9%	45.9%
非词汇化	3.5%	2.6%	6.9%	6.8%	6.1%

表 3 中，F1K 指从训练集中前 1000 句的统计信息。由于 NIST 03 和 NIST 05 句子数目分别为 919 和 1082，所以我们用 F1K 便于公平比较。N03 和 N05 表示在测试集中有用规则的类型分布，可以看出，F1K、N03 和 N05 的分布一致性高。这表明，本文所用方法选择的规则与有用规则具有相同的分布。此外，非词汇化规则只占剪枝规则的 2.6%，而在测试集与 F1K 中的比重超过 6%。这是因为非词汇化规则是最泛化的规则，比其它两种类型规则的频率大。随着语料规模的扩大，会出现比非词汇规则更多的词汇化规则，在这种情况下，非词汇化规则的比重趋于减小。可以看到，剪枝后规则类型的比例发生了显著变化，非词汇化规则所占比例在剪枝后上升，但由于颗粒度较粗，仍然只能占很小的一部分比例，占绝大多数比例的翻译规则还是具有词汇化信息的规则，体现了泛化能力和准确性的一种平衡。

表 4 规则剪枝（基于短语的系统）。注意，由于训练数据和语言模型不同，表 4 中的 BLEU 值不同于表 1 中的 BLEU 值。

	BLEU		模型
	NIST03	NIST05	规模
初始	0.2995	**0.2819**	1.9G
剪枝后(Viterbi)	0.3001	0.2820	**0.12G**
剪枝后(100-best)	**0.3029**	0.2866	0.45G

表 4 为基于短语的 SMT 规则剪枝的实验结果。可以看出，1）基于维特比路径的剪枝方法能减少 95.7%的翻译规则，同时 BLEU 值没有降低；2）基于 100-best 的剪枝方法能减少 76%的翻译规则，且显著性提高（$p<0.05$）；3）剪枝方法在基于句法的系统的性能优于基于短语的系统，主要因为基于句法的 SMT 产生大量的泛化的局部词汇化规则。

相较于基于启发式规则剪枝方法，本文所用方法是基于模型的，实验结果显示对基于句法和基于短语的 SMT 都很有效。

4. 3 模型训练的实验结果

本文在剪枝模型的基础上，使用平均场（mean-field）和半强制解码重新训练模型，主要内容参考 3.2 和 3.3 节。实验结果如表 5 和表 6 所示。

从表 5 和表 6 可以看出，与初始模型与剪枝后的模型相比，重新训练模型后的性能显著提高（$p<0.01$），且模型规模与剪枝后的模型相差不多。充分说明本文所用模型训练方法的有效性。表 5 和表 6 种关于剪枝后的模型，其所使用的概率仍为原始启发式方法中的相对频率，概率估计不准确。当重新训练后，模型所使用的概率为从派生路径中统计出的值，保证了统计量和派生路径一致，最终实验结果显示翻译质量也由于概率估计得更加准确而得到显著提升。

表 5 重新训练模型（基于句法的系统）

	BLEU	
	NIST03	**NIST05**
初始	0.2394	0.2208
剪枝后	0.2486	0.2301
重新训练	0.2532	0.2367

表 6 重新训练模型（基于短语的系统）

	BLEU	
	NIST03	**NIST05**
初始	0.2995	0.2819
剪枝后	0.3001	0.2820
重新训练	0.3042	0.2892

5 总结和展望

本文提出一个通用框架，该框架通过半强制解码和变分贝叶斯 EM 对 SMT 模型进行剪枝和优化。相较于之启发式方法和和基于 EM 算法的框架，该方法在翻译模型上的数学理论基础更强。实验结果显示，该框架对模型的剪枝和优化非常有效。以后的工作将致力于建立更完善的翻译系统，从而降低对启发式方法(Denero 等 2008)的依赖。

参考文献

[1] Alan Agresti. 1996. An Introduction to Categorical Data Analysis. Wiley

[2] C. E. Antoniak. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Annals of Statistics

[3] A. Birch, Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Constraining the phrase-based, joint probability statistical translation model. SSMT Workshop-06. 154–157

[4] D. Blei and M. I. Jordan. 2005. Variational inference for Dirichlet process mixtures. Bayesian Analysis

[5] Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008a. A discriminative latent variable model for statistical machine translation. ACL-HLT-08. 200–208

[6] Phil Blunsom, and Miles Osborne. 2008b. Probabilistic Inference for Machine Translation. EMNLP-08. 215–223

[7] Phil Blunsom, Trevor Cohn, Chris Dyer and Miles Osborne. 2009. A Gibbs Sampler for Phrasal Synchronous Grammar Induction. ACL-IJCNLP-09

[8] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2):263-311

[9] David Chiang. 2005. A hierarchical phrase-based model for SMT. ACL-05. 263-270.

[10] Trevor Cohn and Phil Blunsom. 2009. A Bayesian Model of Syntax-Directed Tree to String Grammar Induction. EMNLP-09. 352-361

[11] A.P. Dempster, N.M. Laird and D.B.Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc., Ser. B. Vol. 39, 138

[12] John DeNero, Dan Gillick, James Zhang, and Dan Klein. 2006. Why Generative Phrase Models Underperform Surface Heuristics. SSMT Workshop-06. 31–38

[13] John DeNero, Alexandre Buchard-Côté, and Dan Klein. 2008. Sampling Alignment Structure under a Bayesian Translation Model. EMNLP-08. 314–323

[14] Matthias Eck, Stephan Vogel and Alex Waibel. 2007. Translation Model Pruning via Usage Statistics for Statistical Machine Translation. NAACL-HLT-07. 21-24 (Companion Volume)

[15] Nicola Ehling, Richard Zens, and Hermann Ney. 2007. Minimum bayes risk decoding for BLEU. ACL-07. 101–104

[16] Jesus-Andres Ferrer and Alfons Juan. 2009. A phrase-based hidden semi-markov approach to machine translation. EAMT-09

[17] T. S. Ferguson. 1973. A Bayesian Analysis of Some Non-parametric Problem. Annals of Statistics

[18] Katerina T. Frantzi and Sophia Ananiadou. 1996. Extracting nested collocations. COLING-96, 41–46

[19] Michel Galley, Mark Hopkins, Kevin Knight and Daniel Marcu. 2004. What's in a translation rule? HLT-NAACL-04

[20] Michel Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang and I. Thayer. 2006. Scalable Inference and Training of Context-Rich Syntactic Translation Models. COLING-ACL-06. 961-968

[21] Zhongjun He, Yao Meng, Yajuan Lj, Hao Yu, Qun Liu. 2009. Reducing SMT Rule Table with Monolingual Key Phrase. ACL-IJCNLP 2009. 121-1245 (short paper)

[22] Gonzalo Iglesias, Adri`a de Gispert, Eduardo R. Banga, and William Byrne. 2009. Rule filtering by pattern for efficient hierarchical translation. EACL-09. 380–388

[23] Abraham Ittycheriah and Salim Roukos. 2007. Direct translation model 2. HLT-NAACL-07. 57–64

[24] Mark Johnson. 2002. The DOP estimation is biased and inconsistent. Computational Linguistics. 28(1): 71-76

[25] Mark Johnson. 2008. A Brief Introduction to Variational Bayesian Inference. http://cog.brown.edu/~mj/ classes/cg168/slides/VariationalBayes.pdf

[26] Howard Johnson, Joel Martin, George Foster and Rol-and Kuhn. 2007. Improving translation quality by discarding most of the phrase table. EMNLP-CoNLL-2007. 967–97

[27] Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for M-gram language modeling. ICASSP-95, 181-184

[28] Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. ACL-03. 423-430.

[29] Philipp Koehn, Franz J. Och and Daniel Marcu. 2003. Statistical phrase-based translation. HLT-NAACL-03, 127-133

[30] Philipp Koehn. 2004b. Statistical significance tests for machine translation evaluation. EMNLP-04, 388-395

[31] Philipp Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL-07 (poster), 77-180

[32] K. Kurihara and M. Welling and Y. W. Teh. 2007. Collapsed variational Dirichlet process mixture models. IJCAI-07

[33] Percy Liang, Alexandre Buchard-Côté, Dan Klein, and Ben Taskar. 2006. An End-to-End Discriminative Approach to Machine Translation. COLING-ACL-06. 761–768

[34] Percy Liang and Dan Klein. 2007. Structured Bayesian Nonparametric Models with Variational Inference. ACL Tutorial. ACL-2007

[35] Yang Liu, Qun Liu and Shouxun Lin. 2006. Tree-to-String Alignment Template for Statistical Machine Translation. COLING-ACL-06. 609-616

[36] Daniel Marcu and William Wong. 2002. A Phrase-based, Joint Probability Model for Statistical Machine Translation. EMNLP-02, 133-139

[37] Daniel Marcu, W. Wang, A. Echihabi and K. Knight. 2006. SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. EMNLP-06. 44-52.

[38] Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. ACL-HLT-08. 192-199.

[39] Haitao Mi and Liang Huang. 2008. Forest-based Translation Rule Extraction. EMNLP-08. 206-214

[40] Jonathan May and Kevin Knight. 2007. Syntactic Re-Alignment Models for Machine Translation. EMNLP-07

[41] Franz J. Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. ACL-02, 295-302

[42] Franz J. Och. 2003. Minimum error rate training in statistical machine translation. ACL-03, 160-167

[43] Franz J. Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. Computational Linguistics, 30(4):417-449

[44] Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin and Dragomir Radev. 2004. A Smorgasbord of Features for Statistical Machine Translation. HLT-NAACL-04

[45] Kishore Papineni, Salim Roukos, ToddWard and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. ACL-02. 311-318

[46] Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. ICSLP-02. 901-904.

[47] Christoph Tillmann and Tong Zhang. 2007. A block bigram prediction model for statistical machine translation. ACM Transactions Speech Language Processing. 4(3):6.

[48] Zhiyang Wang, Yajuan Lv, Qun Liu and Young-Sook Hwang. 2010. Better Filtration and Augmentation for Hierarchical Phrase-Based Translation Rules. ACL-10, 142-146 (short paper)

[49] Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. 2007. Online large-margin training for statistical machine translation. EMNLP-07. 764–773

[50] Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics, 23(3):377-403

[51] Joern Wuebker, Arne Mauser and Hermann Ney. 2010. Training Phrase Translation Models with Leaving-One-Out. ACL-10. 475-484

[52] Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. ACL-01. 523-530

[53] Hui Zhang, Min Zhang, Haizhou Li, Aiti Aw and Chew Lim Tan. 2009a. Forest-based Tree Sequence to String Translation Model. ACL-IJCNLP-2009

[54] Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, Sheng Li. 2008. A Tree Sequence Alignment-based Tree-to-Tree Translation Model. ACL-HLT-08. 559-567

作者简介：高恩婷，女，讲师，主要研究领域为自然语言处理，Email: entinggao@qq.com；段湘煜，男，副教授，主要研究领域为机器翻译、自然语言处理，Email: xiangyuduan@suda.edu.cn；巢佳媛，女，硕士研究生，主要研究领域为机器翻译，Email：jychao@suda.edu.cn；张民，男，教授，主要研究领域为机器翻译、自然语言处理。Email: minzhang@suda.edu.cn。