北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2014.019

基于翻译日志的统计机器翻译模型剪枝

刘凯[1]　吕雅娟[1]　姜文斌[1]　刘群[1,2,†]

1. 中国科学院大学计算技术研究所,北京 100190; 2. 都柏林城市大学,都柏林爱尔兰; † 通信作者, E-mail: liuqun@ict.ac.cn

摘要　提出一种基于翻译日志的统计机器翻译模型的剪枝方法。该方法利用翻译规则，在翻译日志中的命中频数对机器翻译规则进行过滤，保留当前机器翻译模型所需的最小的规则表。实验表明，该方法能够在仅保留原有模型 1%~3%翻译规则的前提下达到原有模型的翻译效果。

关键词　统计机器翻译;模型剪枝;翻译日志

中图分类号　TP391

Statistical Machine Translation Model Pruning based on Translation Log

LIU Kai[1], LÜ Yajuan[1], JIANG Wenbin[1], LIU Qun[1,2,†]

1.Institute of Computing Technology,University ofChinese Academy of Sciences, Beijing 100190;
2.Dublin City University (DCU), Dublin, Ireland; † Corresponding author, E-mail: liuqun@ict.ac.cn

Abstract　The authors propose a novel translation log based translation rule pruning method, which prunes translation rulesaccording to the translation rule hit counts pairs.Experiment results show thatthe proposed method requires only 1%~3% translation rules without significantly difference compared to the full model.

Key words　statistical machine translation; model pruning; translation log

目前基于统计的机器翻译方法中，所使用的翻译模型一般十分庞大，需要较为大型的设备来提供存储和计算支持。较大的模型意味着较高的翻译服务运行成本，同时也意味着难以将翻译服务部署在移动设备上，在一定程度上制约了机器翻译的应用场景。该问题较有前景的解决方案之一就是对翻译模型进行剪枝，缩减翻译模型的规模。

目前现有的一些模型剪枝方法[1-3]致力于在翻译模型各种参数固定之前对规则进行选择，保留翻译时可能需要的规则，去除不需要的部分。其中一些方法利用规则的频度去除低频的规则[4]，该方法能够在保证一定翻译效果的前提下去除大量的低频规则，但是如果去除过多规则的话，该方法将会导致翻译效果显著下降。而另一些方法则利用一些启发式规则，保留一些符合词法或者句法的规则[5–6]，但是这类方法主要的目的为提升翻译质量而非缩减模型，所以难以有效的去除大量规则。

Quirk等[1]认为只抽取最小的短语片段并不会损害太多翻译质量，从而达到减小翻译模型的目的。Johnson 等[2]的工作利用显著性检验去除一些不太可能使用的规则以缩小翻译规则。文献[3–6]利用判别式分类方法(SVM,最大熵)对规则表进行清理。Zettlemoyer 等[4]直接在抽取规则的阶段中不抽取他们认为较差的规则，在源头上控制规则的质量。本文在整个翻译模型确定后再对规则进行过滤，直接利用模型本身的搜索空间信息对规则表进行过滤。与上述方法相比较，我们的方法更加客观、更加符合翻译解码器需求，并且相对简单。

本文提出一种在翻译模型确定后，利用翻译日志对翻译规则进行选择及过滤的方法。我们假设翻译模型中规则使用的频度分布服从长尾定理(Long-Tail)——少量的规则在实际翻译中被频繁使用，而相当部分规则仅被少量使用。在实际的翻译模型中，由于模型各项的参数已经确定，翻译模型

863 计划(2011AA01A207)和国家关键技术支撑项目(2012BAH39B03)资助
收稿日期: 2013–06–15; 修回日期: 2013–09–17; 网络出版时间: 2013-11-06 11:23

的搜索空间有限且已基本确定, 所以存在相当部分的规则在当前参数下难以被有效利用。如果将这部分在翻译模型搜索空间外的“无用”或是很少使用的规则从当前模型中去除, 应当不会对翻译效果产生太多的影响。

由于翻译模型的搜索空间难以通过直接的方法确定, 所以我们提出一种利用翻译日志间接确定翻译模型搜索空间的近似方法,通过去除搜索空间外或边缘上的翻译规则,达到对翻译模型剪枝的目的。进一步, 我们可以配合其他的剪枝方法实现对模型的进一步剪枝, 或者在尽量保留翻译效果的前提下对翻译模型进行剪枝。

本文在较大规模语料训练的翻译模型上进行了相关实验, 结果表明我们的方法能够在翻译效果没有显著改变的前提下, 只需保留原有规则表 2%左右的规则。进一步的, 配合上其他的规则过滤方法, 可以只保留 1%的规则而效果仅有少量的降低, 或者保留 3%的规则来保留原有翻译系统的效果。

1 翻译日志

下面介绍本文使用的翻译日志以及利用该翻译日志得到规则使用频度统计的方法。其中翻译日志为实际系统中翻译过字符串的集合; 规则使用频度为在翻译日志的过程中, 相应规则在所有 1best 译文中被使用的次数统计。

在实际提供服务的翻译系统中, 所有用户提交的待翻译的字符串均可以在系统中保留相应记录。大量的翻译日志记录可以在一定程度上反映用户对文本翻译需求的分布。如果翻译系统能够满足这些翻译日志的翻译需求, 则大部分的翻译需求应当能够被该系统处理, 同时满足这些日志的翻译需求的翻译规则应当在模型的搜索空间内。

由于翻译系统搜索译文的需要, 同一源端的句子将通过不同翻译规则生成许多不同的候选译文, 同时生成相应译文的 NBest 列表。得到候选译文后, 翻译系统将根据模型参数选出最佳的译文作为最终的翻译返回, 同时其他候选译文作为中间结果将不会被返回。如图 1 所示, 翻译系统使用不同的翻译规则将一个中文句子翻译成英语, 生成了两个不同的候选译文(实际翻译中候选译文数目一般更大, 并且候选译文有可能相同)。例子中, 1best 将被返回而 2best 则作为竞争失败的译文不作处理。在实际应用中, 除了 1best 译文之外, 其他候选译文均是无用的, 这些候选译文即使不生成也不会对 1best 译文生成产生太多影响。因此, 在翻译模型确定的情况下, 所有非 1best 译文的搜索路径均是多余的。同理, 除了 1best 译文所需要的规则外, 其他所有规则在当前翻译过程中也是不必要的。例如图 1 中, 如果去除在生成 1best 译文过程中不会使用到的规则(〈沙龙, Salon〉和〈举行了会谈, hold a meeting〉), 只会导致 2best 译文无法生成, 不会影响 1best 译文的翻译结果。因此, 在本工作中仅统计翻译 1best 所使用规则的频数, 该频数统计能够更精确地反映翻译模型对规则的需求分布。

上述的规则频数统计可以从实际的翻译日志中获得。本文利用从翻译日志获得的规则使用频度统计来确定翻译模型的搜索空间, 从而对翻译规则进行过滤, 缩减翻译模型规模。

2 基于翻译日志的剪枝

我们根据已获得的翻译日志, 依此获得相应的规则使用频数统计,利用该频数统计对翻译规则进行剪枝: 直接根据频数统计的剪枝方法或配合频数统计的其他剪枝方法。

2.1 规则使用频数的剪枝

对于已获得的规则频数统计, 我们根据翻译规则被使用的次数对其进行排序。以图 1 中的规则为例, 我们可以得到规则排序如表 1 所示(假设规则〈X1 与 X2, X1 with X2〉在翻译其他句子 1best 译文时也被使用过一次, 则其总计数为 2, 而其他规

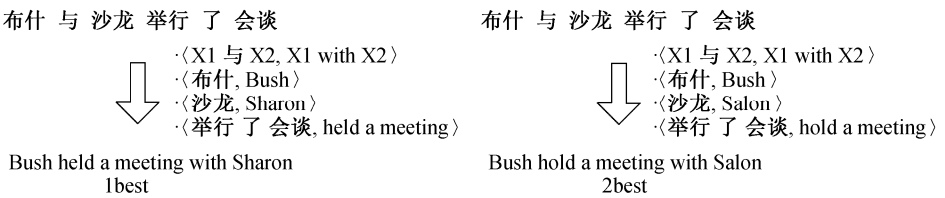


图 1 不同翻译规则获得的两个候选译文

Fig. 1 Two obtained candidate translations through different translation rules

表 1　排序后的翻译规则日志
Table 1　Sorted translation log of rules

规则	规则使用频数
〈X1 与 X2, X1 with X2〉	2
〈布什, Bush〉	1
〈沙龙, Sharon〉	1
〈举行了会谈, held a meeting〉	1
〈沙龙, Salon〉	0
〈举行了会谈, hold a meeting〉	0

则仅在图 1 的例子中被使用)。

显然，翻译规则日志中规则使用计数不为 0 的翻译规则应当都在翻译模型的搜索空间中，而其余计数为 0 的翻译规则很有可能不在翻译模型的搜索空间内。因此，我们假设翻译日志中计数为 0 的规则均在翻译的搜索空间外,可以过滤去除。同理，计数频次较低的翻译规则应当是在翻译的搜索空间的边缘上，只有很小的机会被使用,因此我们可以直接根据日志频次计数对翻译规则进行过滤。以表 1 为例，我们根据不同的翻译频次计数进行翻译日志过滤。

如表 2 所示，我们可以根据不同的频次阈值过滤获得不同大小的规则表。由于该规则表是在当前翻译模型搜索空间上过滤得到的，所以我们认为该规则表可以直接替换当前翻译模型中的原始规则表而不需要重新调整参数。

2.2　联合的翻译日志剪枝

由于翻译模型的搜索空间是近似地由翻译日志中的句子集合获得的，所以规则使用频度统计表示的是规则在当前翻译日志上的分布情况。对于没有出现在翻译日志中的句子集上该分布可能会有较大的偏差，所以我们辅以其他翻译日志无关的剪枝方法对模型进行联合剪枝，保证模型在翻译日志集合外的句子上翻译的质量。

表 2　不同频次对表 1 中的规则进行过滤的结果
Table 2　Filtered result of rules in Table 1

Count=0		Count=1	
保留规则	频数	保留规则	频数
〈X1与 X2, X1 with X2〉	2	〈X1与 X2, X1 with X2〉	2
〈布什, Bush〉	1		
〈沙龙, Sharon〉	1		
〈举行了会谈, held a meeting〉	1		

说明：Count 为过滤规则的频数阈值，小于等于该阈值的规则将不被保留。

表 3　对翻译规则日志根据其他剪枝参数进行重排序
Table 3　Sorted rules according to additional criterion

规则	规则使用频数	其他剪枝参数
〈X1 与 X2, X1 with X2〉	2	3
〈举行了会谈, held a meeting〉	1	4
〈沙龙, Sharon〉	1	2
〈布什, Bush〉	1	1
〈沙龙, Salon〉	0	1
〈举行了会谈, hold a meeting〉	0	0

例如图1中规则“〈沙龙,Salon〉”应当是正确并可利用的，但是在翻译日志时没有被使用过。假设其他的有其他的剪枝方法能够给予该规则适当的分数，我们就可以根据该分数保留部分类似规则。如表 3 所示，我们可以根据其他剪枝方法给相应规则的分数对规则频度表进行重排序，并根据联合参数对规则进行过滤得到最终的翻译规则表。本文使用 3 种较为成熟的辅助剪枝方法作为翻译日志剪枝的补充：基于朴素贝叶斯的剪枝方法；基于模型参数的剪枝方法；基于规则频度的剪枝方法。

2.2.1　基于朴素贝叶斯的剪枝方法

与文献[3]和[6]工作相似，我们将翻译规则过滤作为分类问题进行解决。根据翻译日志的剪枝信息，将翻译规则分成两类：一类在翻译搜索空间内；另一类不在。以此作为训练集，根据以下特征训练朴素贝叶斯分类器：1) 规则正向方向翻译概率及词汇化概率；2) 规则源端及目标端词汇数目；3) 规则源端及目标端词汇数目比；4)规则频度。

得到的朴素贝叶斯分类器可以作为翻译日志方法的一种平滑手段。各规则上由分类器给出的分类概率可以被用作在同样规则日志频度下的一个区分手段。例如表 3 中最后两个规则在日志中的使用频度均为 0，但是沙龙，Salon 我们认为不应当被过滤出规则表。此时如果分类器给其的分数较高，则我们可以根据一定参数将其保留下来。

2.2.2　基于模型参数的剪枝方法

在翻译过程中，翻译解码器在很大的程度上是依据模型的参数和相应规则的分数进行规则选择的，所以翻译时，解码器将根据规则的模型得分、语言模型得分等选出最佳译文。所以，模型分数较高的规则在翻译过程中有较大的概率被选择参与翻译，同理，越高模型得分的规则参与 1best 译文翻译的概率越大。所以利用模型参数和规则的相应分数对规则进行选择应当是一种较为有效的手段。

表 4　利用模型参数和规则相应特征得分获取该规则的模型得分

Table 4　Total score according to the parameters and corresponding weights

	正向翻译概率	正向词汇化分数	反向翻译概率	反向词汇化分数	模型分数
模型参数	1.0	2.0	3.0	4.0	
规则 1	0.1	0.2	0.3	0.4	3.0
规则 2	0.4	0.3	0.2	0.1	2.0

如表 4 所示, 假设我们已知第 2 行的模型特征权重以及下面规则相应的特征得分, 通过内积我们可以获得该条规则在当前翻译模型中的得分。该得分可以作为我们选择模型规则的依据之一。

2.2.3　基于规则频度的剪枝方法

在翻译模型训练抽取翻译规则时可以记录每条翻译规则在训练语料中出现的频度(fractional count), 频度越高, 说明该规则在训练语料中出现的越频繁。我们认为在训练集中频繁出现的规则应当也有较大概率被用作翻译其他语料, 所以该频度在一定程度上可作为过滤翻译规则的依据。我们认为频度越高的翻译规则越可能在翻译过程中使用。

3　实验

本节中我们对基于翻译日志的模型剪枝方法进行验证。首先利用多种语料模拟实际的翻译日志, 并对相应规则频数统计进行分析。在模拟翻译日志的基础上, 分别验证直接的翻译规则日志剪枝和联合翻译日志剪枝的方法。实验中使用 Moses 层次短语作为实验翻译解码器, 短语和规则抽取长度为 7。

3.2　基线翻译模型

本文基线翻译系统利用LDC数据中约150万双语句对的汉英语料库, 来训练汉英翻译模型。语言模型为在法新社语料和新华 Giga 部分语料上训练的 5 元语言模型。在基线模型中, 使用 NIST06 评测数据中的汉英测试集作为基线系统的开发集, 使用 NIST04, NIST05 和 NIST08 作为测试集。

表 5　基线系统在各测试集上的效果

Table 5　Results of baseline system on varying test sets

测试集	BLEU%
NIST04	34.06
NIST05	31.94
NIST08	27.53
测试集均值	31.18

基线系统在各测试集上的效果如表 5 所示。为了模拟实用系统中的配置, 在测试的时候将翻译系统的搜索空间适当的缩小(减小栈的大小等), 以提升翻译的速度。

3.2　翻译日志模拟

为了获取规则频度统计, 首先需要获得相应的翻译日志。在此我们利用 3 种不同的语料模拟现实的翻译日志: 训练语料源端、新闻语料和 web 语料。其中, 训练集源端为训练翻译模型双语语料的源端部分; 新闻语料为与训练语料不重复的汉语新闻领域语料; web 语料使用的是搜狗全网新闻语料库(3 种语料均与测试集不重复)。我们直接利用上述语料获得翻译规则日志, 并统计相应翻译规则日志的信息。

如表 6 所示, 虽然翻译日志语料来源不同, 语料规模相差悬殊, 仍然只有非常小部分的规则被用作 1best 译文的翻译。说明一旦翻译模型确定后, 翻译解码器的实际搜索空间是十分有限的。同时, 翻译规则被使用的次数符合长尾定理。即如图 2 所示, 少量规则被大量使用, 大量的规则仅被少量使用。这证实了我们最初对翻译规则的假设, 说明大量地对翻译规则进行过滤是可行的。

表 6　从不同语料获取的翻译规则频度统计信息

Table 6　Statistical information of varying corpus

翻译日志语料	语料规模	频数>0 规则数目	总翻译规则数目	百分比/%
训练集源端	150w	400w	16704w	2.40
新闻语料	528w	614w	16704w	3.68
Web 语料	606w	465w	16704w	2.78

说明: 频数>0 规则数目为在当前翻译日志统计下, 翻译规则日志频次不为零的规则数目。总翻译规则数目为原模型中的总规则数目。相应的百分比为频次不为 0 的规则数目占原有所有规则数目的百分比。

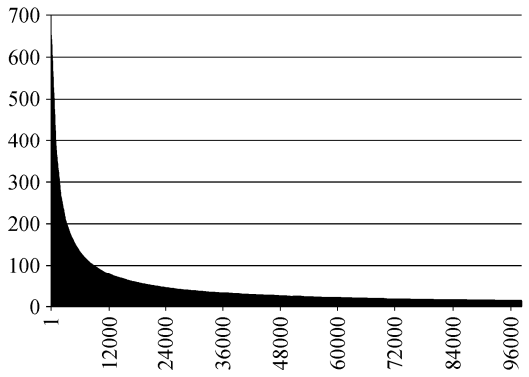


图 2　使用频次前十万规则的使用频次

Fig. 2　Frequency of top 100000 rules

表 7　根据不同语料及过滤频度过滤规则表的测试结果

Table 7　Results of the proposed method

翻译日志	训练集			新闻语料	Web 语料
过滤频度	Count=0	Count=1	Count=2	Count=0	Count=0
规则表大小%	2.40	0.89	0.47	3.68	2.78
NIST04 (BLEU%)	34.01	33.08	32.40	33.74	32.76
NIST05(BLEU%)	31.46	30.78	30.12	31.07	30.58
NIST08(BLEU%)	27.24	26.45	25.80	26.70	26.50
测试集均值	30.90 (−0.28)	30.10 (−1.08)	29.44 (−1.74)	30.50 (−0.68)	29.95 (−1.23)

注: 括号中的值为与基线翻译模型的差距。

3.3　翻译规则日志剪枝

根据上一节实验中获得的翻译规则使用频度, 我们利用由不同语料获得的翻译规则日志以及不同的翻译日志过滤频度对翻译规则进行过滤, 并直接利用过滤后的规则表替换原有模型中的规则表, 在测试集上进行测试。

从表 7 可以看出, 直接利用翻译规则日志中的频度对翻译规则进行过滤, 已经能够在只保留极少量的规则的前提下, 保持与原翻译系统相当的效果。在同样的过滤频度下, 利用训练集源端获得的翻译规则日志进行过滤的翻译效果最好, 并且规则保留的最少。该现象一方面的原因是因为测试集均是与训练集不重复的新闻语料, 利用训练集作为翻译日志保留了最多的词汇翻译信息, 在过滤频度为 0 的时候不会因为过滤规则导致额外的未登入词问题。通过实验结果还可以看出, 根据该方法过滤后的规则表无需重新调整参数也能够保持良好的翻译效果。进一步地, 我们使用效果较好的两组过滤后的规则表(训练集和新闻语料根据 Count=0 过滤的规则表), 重新进行最小错误率训练(MERT), 调整模型参数。

如表 8 所示, 利用过滤后的规则表重新进行参数调整后的效果要略微好于直接使用原始参数的效果。但是效果并不显著, 所以根据实际需要我们可以无需重新调整模型参数。并且根据不同翻译日志语料的效果得出一个结论: 训练集源端是一个合适的翻译日志选择, 能在很大程度上保持翻译效果良好且只需保留最少的规则数目。

表 8　利用过滤后的规则表重新进行模型参数调整后的结果

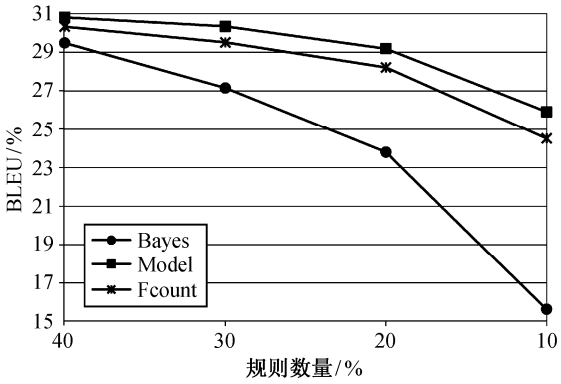
Table 8　Results of re-tuning

翻译日志	训练集	新闻语料
测试集均值(重调参前)	30.90	30.50
测试集均值(重调参后)	31.03	30.90

3.4　联合的翻译日志剪枝

利用上一节获得的翻译规则日志并配合其他的规则过滤方法, 可以更灵活地对翻译规则进行过滤。首先我们利用其他不同的规则过滤方法对翻译规则进行过滤, 并在测试集上进行测试,结果如图 3 所示。

可以看出 3 种辅助的规则过滤方法在保留 20%以下的规则数目时翻译效果急剧下降。但是如果以这些过滤方法作为翻译日志剪枝方法的补充, 则可以更灵活的对翻译规则进行过滤。我们利用这些方法给出的分数作为翻译规则使用频数的补充, 当规则频数相同时, 我们选择补充方法分数较高的规则。我们利用该方法, 将规则表过滤至 1%~9%不等大小, 并测试相应规则表的翻译效果。

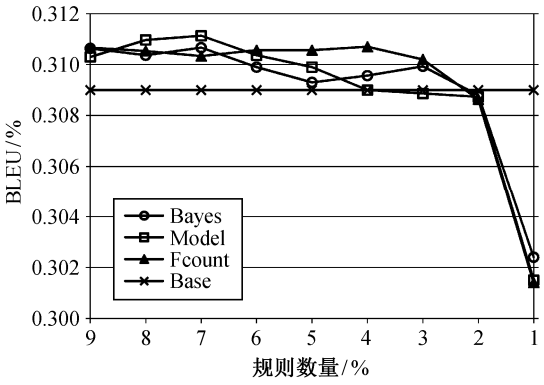


Bayes: 朴素贝叶斯剪枝方法; Model: 模型参数剪枝方法; Fcount: 规则频度的剪枝方法

图 3　不同过滤方法将翻译规则过滤到相应数量时的翻译效果

Fig. 3　Performance of different pruning methods on varying sizes of rules

如图 4 所示, 配合翻译规则频度的方法的翻译日志方法过滤效果更为稳定。且在保留 3%左右翻译规则时, 翻译效果与未过滤翻译规则的翻译模型相



Base: 基本的翻译日志剪枝结果(日志语料: 训练集, Count=0)

图 4　不同过滤方法配合下的翻译日志剪枝方法

Fig. 4　Performance of different pruning methods

近(无显著差别)。同时由于加入了其他过滤的标准, 使得我们可以更灵活地选择保留翻译规则数目。

4　结论

基于翻译模型搜索空间有限且翻译规则使用的频数符合长尾定理的假设, 本文提出了一种基于翻译日志的翻译模型剪枝方法, 并考察了该方法与其他剪枝方法联合剪枝的策略。该方法可以仅利用翻译训练语料源端, 即可实现对翻译模型中的翻译规则的大规模剪枝。实验表明, 我们的剪枝方法只需保留原有规则表 1%~3%的规则即可实现与原有模型无显著差别的翻译效果。我们未来的工作是利用相似的方法对翻译模型中的语言模型进行剪枝, 从而实现对整个翻译模型进行剪枝, 最终降低翻译服务的运行成本, 使得翻译系统在移动设备上的部署成为可能。

参考文献

[1] QuirkC, MenezesA. Do we need phrases?: challenging the conventional wisdom in statistical machine translation // Proceedings of the main conference on human language technology conference of the North American chapter of the Association of Computational Linguistics(NAACL). New York City, USA: Association for Computational Linguistics 2006: 9–16

[2] JohnsonH, MartinJ, Foster G,et al.Improving translation quality by discarding most of the phrasetable // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).Prague, Czech Republic: Association for Computational Linguistics, 2007: 967–975

[3] KavithaKM, Gomes L, Lopes G P, et al.Using SVMs for filtering translation tables for parallel corpora alignment// EPIA, 2011: 690–702

[4] ZettlemoyerLS, Moore R C.Selective phrase pair extraction for improved statistical machine translation // Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics. Rochester, New York: Association for Computational Linguistics, 2007: 209–212

[5] Tu Zhaopeng, Liu Qun, Lin Shouxun. Extracting long distance reordering rules with dependency restriction. Journal of Chinese Informaiton Processing, 2011, 25(2): 55–60

[6] Liu Qun, He Zhongjun, Liu Yang, et al. Maximum entropy based rule selection model for syntax-based statistical machine translation // Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP). Honolulu, Hawaii: Association for Computational Linguistics, 2008: 85–974