翻译规则优化中的分层优化方法

刘树杰，杨沐昀，赵铁军

（哈尔滨工业大学计算机学院机器翻译实验室，哈尔滨，150001）

摘　要： 在基于规则的机器翻译系统中，对规则库进行优化能够删除冗余和冲突的规则，提高规则库的质量。然而规则优化的过程如果处理不当会删掉一些好的规则。如何避免这种情况的发生，是规则优化过程中很重要的问题。本文提出了通过在规则优化的过程中采用分层优化的思想，避免不同层次的规则进行竞争，来避免上述情况的发生。实验证明能够取得比较好的效果。

关键词： 机器翻译； 规则优化； 句法树；

A CASCADED APPROACH TO THE OPTIMIZATION OF TRANSLATION RULES

SHUJIE LIU, MUYUN YANG, TIEJUN ZHAO

(MOE-MS Key Laboratory of Natural Language Processing and Speech, Harbin Institute of Technology, Harbin, 150001)

Abstract: As far as the rule-based machine translation (RBMT) is concerned, the rule acquisition remains as a bottle-neck problem. This paper proposes a cascaded approach to optimization of the rule base automatically acquired from the bilingual corpus. Observing the more risk of errors in the upper layer of the parsing tree, this method advocates the optimization of rules by a bottom-up strategy so as to take the advantage of correctness of parsing results near the leaf nodes. The experimental results further prove that such cascaded optimization out-performs the usual practice.

Keywords: Machine Translation; Optimization of Rules; Syntactic Tree;

1　绪论

机器翻译是利用计算机把一种自然语言转变成另一种自然语言的过程。现在的机器翻译方法分为基于规则（理性主义的方法）的机器翻译和基于语料（经验主义的方法）的机器翻译两类，基于语料的机器翻译又有基于实例的机器翻译和基于统计的机器翻译等。

基于规则的机器翻译一直是机器翻译的主流方法。基于规则的机器翻译是对源语言和目标语言的词法，句法进行分析，形成源语言到目标语言的翻译规则，然后利用规则对源语言进行处理，形成目标语言。规则的优劣从根本上决定着翻译质量的好坏。

为了提高译文的质量，需要不断的扩充规则库来处理不同的情况。然而新添加的规则可能与规则库中原本存

基金资助：国家自然科学基金（项目号：60375019）

作者简介：刘树杰（1982-），男，山东潍坊，硕士在读　E-mail:shujieliu@mtlab.hit.edu.cn

在的规则存在冲突和冗余。于是人们提出了很多关于规则优化的方法来改进规则库：(1)Menezes 提出了使用频度过滤的方法来删除低频匹配规则[6]，但该方法仅能提高翻译系统的效率，并不能从根本上提高输出译文的质量。(2)Lavoie[7]首次使用对数似然比的方法来对候选规则进行排序，同时使用了错误驱动的策略来消除影响译文质量的规则。(3)Imamura[5] 使用 x2 测试的方法来获取抽象规则。但这些方法仅限于处理高频规则。随着自动译文评测技术的快速发展，人们开始使用 BLEU[8]和 NIST[9]方法对机器译文进行自动评价[10,5]。(4)Kenji[11] 以自动译文评测分数增长为依据来删除规则库中影响译文质量的规则，使得机器译文的质量有了一定程度的提高。Kenji 的方法在对规则优化的过程中没有考虑到规则的层次性对规则优化过程的影响。对规则进行优化时会存在不同层次的规则相互影响的情况出现，这种情况会造成一些好的底层的规则由于高层的不好规则而被删掉。如果要避免这种情况的发生，在规则优化的时候就应该考虑不同层次规则的竞争。即，将规则优化的过程按层次进行。

优化的过程按层次进行的基本思想：基于规则的翻译过程是在句法树上进行规则匹配的过程，如果对规则优化的过程采取逐层添加、逐层优化的方法，尽量让同层的规则进行竞争，避免不同层次的规则的干扰，能够改进优化效果。

2 规则的树形结构和分层优化的思想

2.1 规则的树形结构及句法树的翻译过程

基于规则的机器翻译有两种: Tree-to-Tree 和 Tree-to-String。Tree-to-Tree 的方法受目标语言和源语言两种语法体系的制约,因而效果不是非常理想。而 Tree-to-String 的方法由 Fai WONG 提出并经证明是比 Tree-to-Tree 较优的方法。我们这儿使用的是 Tree-to-String。

Tree-to-String 的规则抽取过程[1]:

1.对双语句对中的英语句子进行词型还原;

2.使用汉语句法分析器对双语句对中的汉语句子进行句法分析，获得汉语句法树;

3.使用词对齐工具对双语句对进行词对齐;

4.依据词对齐信息获得汉语句法分析树中每一个非叶结点对应的英语译文片断;

基于 Tree-to-String 规则的机器翻译过程：首先将原语言的输入句子经过句法分析形成句法树，然后从规则库中选取匹配的规则对句法树进行处理，形成目标语言的句子。

基于 Tree-to-String 规则的机器翻译往往经过三个阶段：词汇级转换、短语级转换和句子级转换。

词汇级转换： 该过程往往是借助词典完成词义消歧和译文选择。

短语级转换： 通过规则进行转换。

句子级转换： 句子级的调整过程。解决的是两种语言在句子模式一级的差异。

转换的处理过程是：自底向上的，在遇到一个短语符号时，立即进入该短语对应的知识库，进行节点合并、转换、生成以及特征传递工作；直至最后进入句子级的转换。

举例：哈尔滨工业大学计算机学院机器翻译实验室 Cemt2003 的汉英翻译流程如图所示。

2.2 分层优化的思想

Kenji 采用的优化方法是总体优化的方法,该方法没有考虑到规则的层次性,其过程为:采用逐步删除策略。在初始时将所有候选规则加入系统规则库中。然后，对规则库中的每条规则，首先从规则库中删掉此规则，用删掉该规则后的规则库翻译训练语料，并对结果进行评分，如果评分变好，则删掉该规则，否则保留。对规则库中的所有规则执行此过程，直到所有规则都遍历一遍为止。这样的优化方法没有考虑不同层次间规则的相互影响。

虽然规则没有绝对的层次性,比如规则 A 可能在句子 1 的句法树中出现在最低层而在句子 2 中出现在最高层。然而对于大多数规则有着出现在固定层次的偏好。比如：对于一个形容词短语（BAP[最/d 高/a]）抽取到的规则：0:Cate=d+0:W=最+1:Cate=a+1:W=高->0:。该规则很明显偏向于出现在句法树的底层。即规则之间存在不很绝对的层次性(该现象可以通过对规则的层次进行统计得出)。

考虑到规则的层次性，在规则优化的过程中会出现这样的情况，对于一条比较好的底层规则 A，当在用删掉该规则后的规则库进行翻译时会因为一条比较差的高层规则 B 的影响而导致删掉 A 后的规则库翻译的效果质量更

好。然而由于规则 B 是比较差的规则，所以在以后的处理过程中极有会被删掉，这样就会出现因为一条高层的差的规则导致底层的比较好的规则被删掉的情况出现。这种现象的出现是因为在优化时出现了不同层次的规则 A、B 进行竞争的情况。

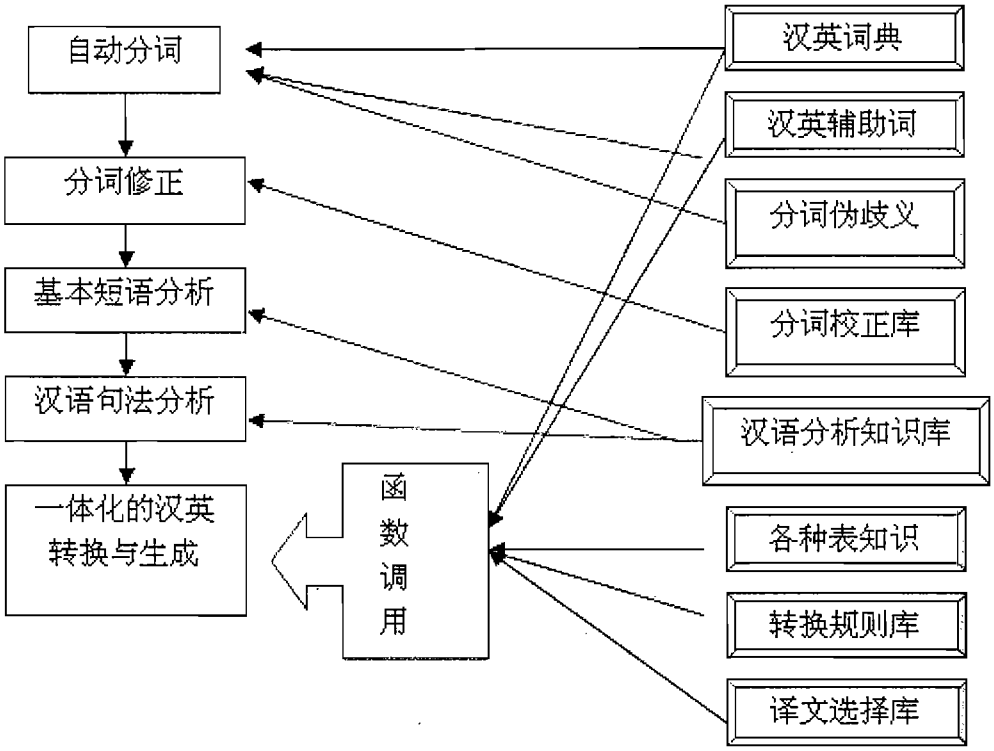

图 1 哈尔滨工业大学计算机学院机器翻译实验室 Cemt2003 的汉英翻译流程

Fig.1 The process of the Cemt2003 of MOE-MS Key Laboratory of Natural Language Processing and Speech (Harbin Institute of Technology)

我们试图寻找一种方法能够最大程度的避免这种情况(即:避免不同层次的规则的相互干扰)的发生。我们认为的最好的优化过程是：

1， 每条规则都有层次标记(对规则引入层次属性)，按规则的层次标记添加规则，并进行优化。

2， 我们假设经过优化的所有层的规则都是好的，不存在冲突和冗余，所以在优化的时候只是遍历刚刚添加进来的规则，并判断该规则是否应该保留，这样就能够避免不同层次规则的竞争。

这种方法的缺点是：给规则添加层次标记，就会限制规则的适用范围。从而会出现很多除了层次不一样其余都一样的规则，降低了召回率，虽然提高了精确率。

然而，基于时间和工作量的考虑没有采取这种方法，而是从已经抽取的规则中取了一个样本（10000/610778）统计各种类型规则的平均层数（统计结果见附录），然后根据统计结果对规则按句法标注类型进行排序，按照排好的句法标注类型顺序进行优化。并且在删掉规则的时候也没有考虑到是否是出自刚刚添加的这一层。

3 实验过程

本实验优化过程中所用的翻译系统是哈尔滨工业大学机器翻译实验室的 cemt2003 的翻译模块,优化过程中的评测方法是 bleu 评测。实验从 30 万双语语料中抽出了 230900 条嵌套规则和 489237 条嵌套规则。规则优化过程使用的汉语树库为 610778 句及相应的英语译文。该 30 万双语语料库是内容涉及对话、旅游、法律、艺术、生活等各个方面的词典例句语料库。测试语料是 863 语料库的对话部分。

首先将 30 万句对的语料经过词汇对齐、句法分析、等价对获取、规则抽取等步骤处理，获得 230900 条非嵌套规则和 489237 条嵌套规则。从等价对抽取的 610778 等价对中选择 10000 个局对进行层次统计。层次统计的样本是类似如下的等价对:AP[BMP[一/27/m 条/28/q]狭窄/29/a]的/30/usde-->a/19 narrow/20。有一条嵌套的

规则从该等价对抽取出来并且该规则的层数定义为：3，计算每个等价对的层数，然后统计。得到统计结果，按照统计结果对句法标注类型排序。

我们在安排类型的顺序的时候首先将类型按处于第一层的规则数所占的比例从大到小进行排序，抽取前 6 个类型，然后按照处于第二层的规则数所占的比例从大到小进行排序抽取 10 个，然后第三层抽取 10 个，其余的按最大层数最小的排序。排完序后进行了手工的调整。

统计结果举例如下：

```
#AP
1:769,  prop:0.618665   2:265,  prop:0.213194   3:112,  prop:0.090105
4:52,   prop:0.041834   5:28,   prop:0.022526   6:11,   prop:0.008850
7:3,    prop:0.002414   8:3,    prop:0.002414
```

说明：#AP　1 :769, 占总数的:0.618665

表示所有的 AP 类型的规则中有 769 条规则是从第一层抽取的，占总数(769+265+112+52+28+11+3+3) 的 0.618665。

对非嵌套规则进行总体优化。由于非嵌套规则都是位于句子的底层的规则，所以不需要进行分层优化处理。

在优化后的非嵌套规则的基础上添加排好序的嵌套规则的第一个类型的规则，进行优化。优化后依次添加第二种类型的规则，进行优化（优化的算法采用逐步删除策略，如前），如此循环，直到添加完所有的规则。

规则添加的顺序为：BAP BMP BDP BVP BNT BNS MP VSVO BNP VBA CO NP NT ASIDE AP DP NDE NS PP INP PFP VBEI VC VJ VO VOO VP VV XP SS SENT。 规则符号类型说明见附录。

规则优化的评价函数为[2]：

$$Score(r)=\left[\left(\frac{n_{pos}+1}{n_{neg}+1}\right)^{-\mathrm{sgn}(s)}-\theta\right]*\mathrm{sgn}(s) \qquad (1)$$

$$\mathrm{sgn}(s)=\begin{cases}-1 & s<=0\\ 1 & s>0\end{cases} \qquad (2)$$

sgn(s)： 符号函数-当 s 为正时其值为 1，否则其值是-1；

θ：在评价规则时，用于平衡规则对译文评测分数的贡献与规则正确转换语言现象能力的一个因子，其值为 1.0。

4 实验结果及分析

表 1 试验结果

Tab.1 The results of experiment.

评分方法	分层优化	不分层优化
bleu1	0.607099441	0.607881132
bleu3	0.177000925	0.17665358
bleu5	0.097127068	0.097009652

实验结果表明分层优化的结果略好于不分层优化的结果。然而结果并不是非常的明显，是因为实验的过程中存在一些缺陷。主要表现在：规则的添加顺序并不是按照规则抽取时所在的层数添加的，而是按 1/30 的规则统计出来的不同类型大体上处于什么层次来进行添加的，所以这样的结果与理想的情况肯定会存在差距。实验的过程还存在另外一些原因导致实验没能按照理想的步骤进行。由于时间等各方面的原因，没有再做额外的实验来证明结果具有统计意义。

如果采用前边介绍的最理想的方法进行实验应该能够取得更好的结果。

参考文献：

[1] Muyun Yang[, Zhanyi Liu, Tiejun Zhao]. TBED Based Chinese-English Translation Rule Acquisition.Proc. of IEEE Conference on Natural Language Processing and Knowledge Engineering. Beijing, China, Oct. 2003, 26-29

[2] 张春祥[, 李生, 杨沐昀, 赵铁军， 时晓升]. 一种基于评价的规则库优化方法 哈尔滨工业大学报(录用但未发表).

[3] 齐浩亮[, 杨沐昀][,等]. 面向特定领域的汉语句法主干分析。 中文信息学报 2004

[4] Setsuo Yamada[, Kenji Imamura, and Kazuhide Yamamoto]. Corpus-Assisted Expansion of Manual MT Knowledge. The 9th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002), 2002:199-208

[5] Kenji Imamura. Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment for Pattern-based MT. In the 9th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002), 2002: 74-84

[6] Arul Menezes[, Stephen D. Richardson]. A best first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In Proceedings of the 'Workshop on Example-Based Machine Translation' in MT Summit VIII, 2001:35-42

[7] Benoit Lavoie[, Michael White, Tanya Korelsky]. Inducing Lexico-Structural Transfer Rules from Parsed Bitexts. In Proceedings of the ACL 2001 Workshop on Data-driven Machine Translation, 2001:17-24

[8] Kishore Papineni[, Salim Roukos, Todd Ward et al]. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics(ACL), 2002:311-318.

[9] George Doddington. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In Proceedings of ARPA Workshop on Human Language Technology, 2002

[10] Sonja Nien[, Franz J. Och, G. Leusch et al]. An evaluation tool for machine translation: Fast evaluation for machine translation research. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC), 2000:39-45

[11] Kenji Imamura[, Eiichiro Sumita, Yuji Matsumoto]. Feedback Cleaning of Machine Translation Rules Using Automatic Evaluation. In 41st Annual Meeting of the Association for Computational Linguistics, 2003: 447-454.

附录：

规则符号类型说明：

共 31 种符号，各符号说明如下：

BAP	基本形容词短语	CO	并列结构	PFP	方位结构
BMP	基本数量短语	NP	名词短语	VBEI	被字结构
BDP	基本副词短语	NT	时间名词短语	VC	动补结构
BVP	基本动词短语	ASIDE	似的结构	VJ	兼语结构
BNT	基本时间短语	AP	形容词短语	VO	动宾结构
BNS	基本处所名词短语	DP	副词短语	VOO	双宾结构
MP	数量短语	NDE	的字结构	VP	动词短语
VSVO	所字结构	NS	处所名词短语	VV	连动结构
BNP	基本名词短语	PP	介词短语	XP	搭配结构
VBA	把字结构	INP	插入语或独立成分	SS	小句或子句
SENT	句子				