统计机器翻译中的源语言重排序方法研究*

梁芳丽[1,2]，李 淼[1]，李 文[1,2]，陈 雷[1]，乌达巴拉[1]

[1]中国科学院 合肥智能机械研究所，安徽 合肥 230031

[2]中国科学技术大学 信息科学技术学院，安徽 合肥 230026

E-mail: lfli8609@mail.ustc.edu.cn

摘 要：为了更好地解决统计机器翻译中的调序问题，本文提出了基于句法信息、词性标注信息和规则相结合的源语言重排序模型作为统计机器翻译的预处理模块。该模型分为两种，一种是基于依存信息、词性标注信息和规则相结合的模型，另一种是基于短语结构信息、词性标注信息和规则相结合的模型。以汉蒙统计机器翻译做实验，结果显示经过该模型进行预处理后的统计机器翻译的BLEU评分比经典的短语翻译有较为明显地提高。实验结果表明这两种源语言重排序模型都是有效的，都能较为显著地提高译文的质量。

关键词：统计机器翻译；依存信息；词性标注信息；短语结构信息；源语言重排序

Research on Methods of Source Language Reordering in Statistical Machine Translation

Liang Fangli[1,2], Li Miao[1], Li Wen[1,2], Chen Lei[1], Wudabala[1]

[1] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031

[2] School of Information Science and Technology, University of Science and Technology of China, Hefei 230026

E-mail: lfli8609@mail.ustc.edu.cn

Abstract: In order to solve the reordering problem of Statistical Machine Translation (SMT), a source language reordering model based on syntactic information, POS tagging information and rules is proposed as a preprocessing module. The model is divided into two. One is based on the combination of dependency information, POS tagging information and rules, the other is based on the combination of phrase structure information, POS tagging information and rules. Take the example of Chinese-Mongolian SMT, experimental results show that the BLEU scores of using the model as preprocessing are more obviously improved than a conventional phrase-based SMT. The results imply both of source language reordering models are effective and can improve the quality of translation more significantly.

Keywords: statistical machine translation; dependency information; POS tagging information; phrase structure information; source language reordering

1 前言

在统计机器翻译系统中，源语言和目标语言的语序往往存在较大的差异，因此调序在机器翻译中占据着关键的地位。对统计机器翻译而言，其调序模型可分为基于特征融入的对数线性模型和基于语序调整的预处理模型。前者易于将有用的知识源以特征的形式添加到模型中，后者是在翻译之前对源语言句子进行调整使之和目标语言的词序更接近。第一种调序模型[1-3]在训练和解码时需要大量的额外工作和时间耗费，相比较而言，第二种调序模型在训练和解码时有降低时间开销的明显优势。在这类调序模型中，有基于源语言的重写模式模型[4]、基于从句转换规则的模型[5]和融入句型信息的调序模型[6]。

近年来，基于句法和规则相结合的源语言重排序模型[7]逐渐被应用。Nizar Habash [8]提出了基于源语言依存树和单词对齐自动抽取重排序规则调整源语言的方法；Xu Peng [9]等人提出了基于谓语头是动词、形容词、名词或介词驱动的带有优先次序的重排序规则进行源语言调序的方法；孙

* 基金项目：国家自然科学基金(61070099)，国家科技支撑计划(2009BAH41B06)

广范[10]等人介绍了基于汉语中“的”字引导的相关短语的句法调序方法；Karthik Visweswariah[11]描述了基于源语言短语结构树和单词对齐自动抽取重排序规则进行源语言调整的方法；Young-Suk Lee[12]等人提出了基于上下文无关文法和上下文敏感文法的重排序规则调整源语言语序的方法。然而对于这些句法和规则相结合的方法，会出现由句法分析错误导致的抽取规则的不准确性，书写规则的片面性以及规则和规则之间会产生相互影响等不足。

为了降低由这些不足造成译文质量下降的影响，本文提出了使用句法局部信息和加入词性标注信息的方法，即使用基于句法信息、词性标注信息和规则相结合的源语言重排序模型的方法。本文研究了两种不同的源语言重排序模型，一种是基于依存信息和词性标注信息并结合规则的重排序模型，另一种是基于短语结构信息和词性标注信息形成重排序规则的模型。这两种不同的源语言重排序模型在一定程度上弥补了句法和规则相结合的不足。实验结果表明，将词性标注信息、句法信息和规则进行结合，能够增强统计机器翻译的调序能力，很好地提高译文的质量。

2 依存信息和词性标注信息相结合的源语言重排序

基于依存信息和规则的源语言重排序的思想是利用依存分析得出的词和词之间的依赖约束关系抽取出依存重排序规则，然后基于这些重排序规则在依存树上进行源语言的语序调整。

图 1 给出了一个汉语句子所对应的依存树。在图 1 中，“ROOT”是依存树的根节点，对应的词“需要”在句子中作谓语，“dobj”是根节点 ROOT 下的一个孩子节点，对应的词“钱”在句子中作直接宾语。

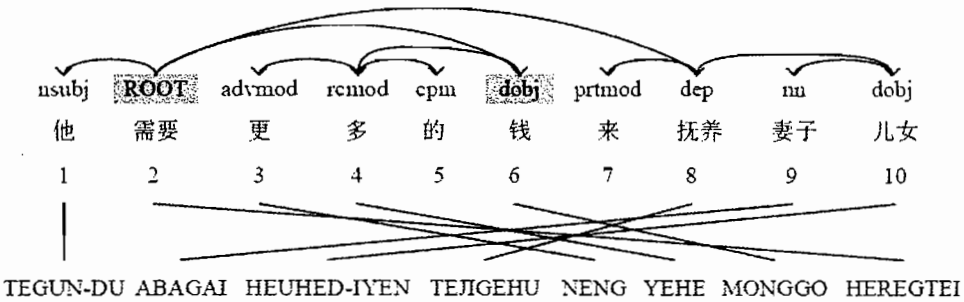

图 1 一个汉语句子的依存分析树

依据依存重排序规则[8-9]，很容易将“需要 更 多 的 钱”调整为“更 多 的 钱 需要”。调序后的汉语“更 多 的 钱 需要”更符合蒙古语“NENG YEHE MONGGO HEREGTEI”的语序。在该图中的汉语“抚养 妻子 儿女”也需要语序的调整，但是由于没有对应的重排序规则，因而该方法不能够捕捉到这样的调序信息。为此，我们提出了在此基础上加入词性标注序列信息的方法，即基于依存信息、词性标注信息和规则相结合的源语言重排序方法。

基于依存信息和词性标注信息并结合规则进行源语言调序的基本思想是首先由依存信息得到谓语和直接宾语，然后将谓语和直接宾语作为边界，找出它们之间的词性标注序列，形成有效的词性标注序列模板，最后采用最大匹配模板法进行源语言语序的调整。

对于词性标注序列模板的形成，首先根据由依存分析得到的谓语和直接宾语及其对应的位置信息形成谓语-位置信息和宾语-位置信息的序列：

$S_{rd}=\langle r_{i1}-rp_{i1}, r_{i2}-rp_{i2}, \cdots, d_{i1}-dp_{i1}-dp_{i2}, \cdots\rangle$，其中 r_{ik}、rp_{ik}、d_{ik} 和 dp_{ik} 分别代表第 i 个句子中的谓语、谓语所在的位置、宾语和宾语所在的位置。

其次，对这些序列按位置信息进行排序，如果排序后的序列中出现了子序列 $\langle r_{is}-rp_{is}, d_{it}-dp_{it}\rangle$，则表示 (r_{is}, d_{it}) 是有效的谓语宾语对，此时记录对应的位置信息 (rp_{is}, dp_{it})。

然后在句子对应的词性标注序列$P_1,P_2,\cdots,P_n$中检查rp_{is}到dp_{it}之间有没有表示标点符号的 PU 出现。若没有，则记录它们之间的词性标注序列$P_{rp_{is}},P_{rp_{is+1}},\cdots,P_{dp_{it}}$。接着，遍历词性标注序列模板，如果模板中不存在该序列，则将该序列加入到该模板中，如果存在则不加入。重复以上过程，最终形成一个包含不同词性标注序列的模板。

在形成有效的词性标注序列模板后，首先采用最大匹配模板法找出源语言句子中匹配的词性标注序列，并记录对应序列的首尾处的位置 iu 和 iv。对一个句子所对应的任何二个位置信息对(iu,iv)和(ix,iy)进行过滤，去掉出现交叉(iu<ix≤iv<iy)和包含(iu≤ix<iy≤iv)关系的位置信息对。根据过滤后的有效位置信息对(iw,iz)，对 iw 和 iz 处的单词进行交换，从而完成源语言的重排序。

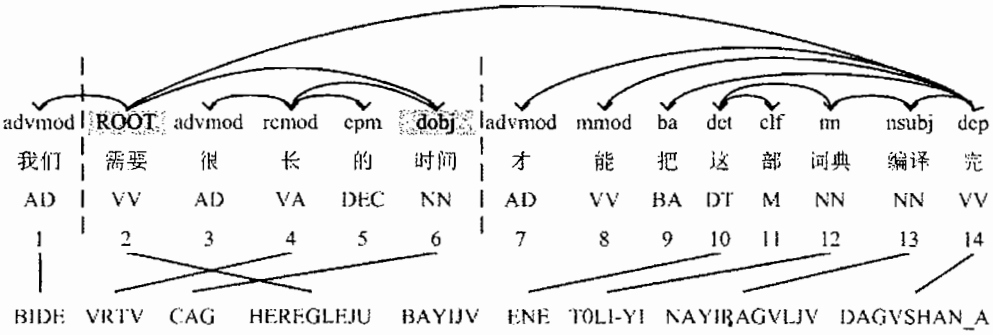

图2 一个汉语句子对应的依存树和词性标注

他	需要	更	多	的	钱	来	抚养	妻子	儿女
PN	VV	AD	VA	DEC	NN	MSP	VV	NN	NN
1	2	3	4	5	6	7	8	9	10

图3 一个汉语句子对应的词性标注

图 2 给出了一个汉语句子的依存树和词性标注。在抽取有效词性标注序列模板时，首先找到该句子的谓语“需要”和直接宾语“时间”，并记录其对应的位置信息 2 和 6；在原词性标注序列中，检查从第 2 个位置到第 6 个位置之间的词性标注序列“VV AD VA DEC NN”，发现没有表示标点符号的 PU 出现，则将未在模板中存在的词性标注序列“VV AD VA DEC NN”加入到模板中。

针对图 1 中的句子，其对应的词性标注信息如图 3 所示。在应用词性标注序列模板进行源语言重排序时，首先找出符合模板的最长的词性标注序列“VV AD VA DEC NN”和“VV NN NN”及其对应的起止位置信息(2,6)和(8,10)；其次利用位置信息(2,6)和(8,10)分别将“需要 更 多 的 钱”调整为“更 多 的 钱 需要”，“抚养 妻子 儿女”调整为“妻子 儿女 抚养”。从该例子可以看出，该方法通过词性标注序列“VV NN NN”捕捉到了“抚养 妻子 儿女”的调序信息。

3 短语结构信息和词性标注信息相结合的源语言重排序

基于短语结构信息和词性标注信息相结合的源语言重排序的基本思想是首先根据源语言和目标语言出现语序差异的部分，整理出带短语结构信息和词性标注信息的重排序规则，然后将这些重排序规则应用到短语结构子树上从而实现源语言的语序调整。

3.1 短语结构信息和词性标注信息相结合的重排序规则

重排序规则是结合巴达玛敖德斯尔所著的《面向机器翻译的汉蒙短语转换规则研究》一书所提出的转换规则进行整理得出的。这些规则中不仅包括短语结构标签，而且还融入了词性标注标

签，如表 1 所示。

表 1　重排序规则

序号	原规则	重排序规则
1	VP→VP_1 VP_2	VP→VP_2 VP_1
2	VP→VV PP	VP→PP VV
3	VP→VV NP	VP→NP VV
4	VP→VV QP	VP→QP VV
5	PP→P NP	PP→NP P

在表 1 中，1~4 表示根节点是动词短语的重排序规则，1 代表的是动词短语与动词短语之间的调序规则，2、3 和 4 分别代表的是动词与介词短语、名词短语和量词短语之间的调序规则，5 代表的是根节点是介词短语的介词与名词短语之间的调序规则。这些规则在很大程度上反映了汉语和蒙古语之间出现差异的部分。

3.2　基于重排序规则和短语结构子树的源语言重排序

首先，由于源语言句法分析后得到的短语结构树是括号短语结构树，这种结构不方便重排序规则的应用，所以通过括号转换算法将括号短语结构树进行结构形式的转换，转换的结果如图 4 所示。

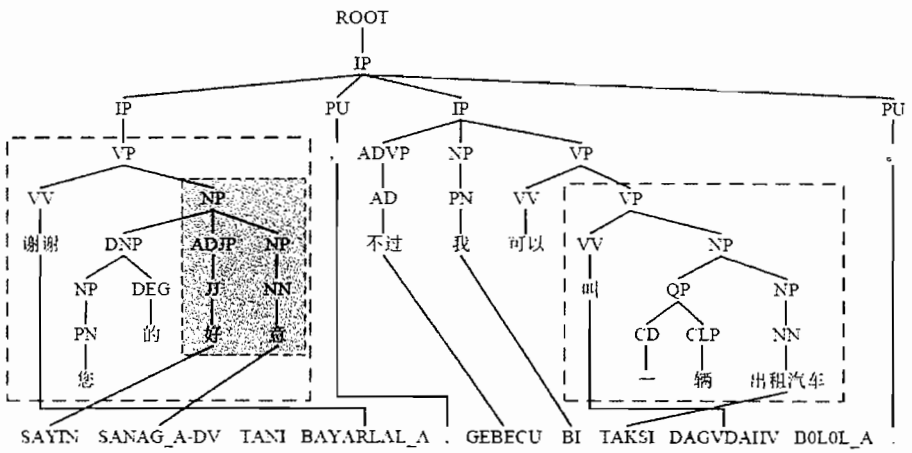

图 4　转换过的短语结构树

其次，将重排序规则应用到转换过的短语结构子树上，具体过程如算法 1。

算法 1 基于重排序规则和短语结构子树的源语言重排序算法

输入：源语言短语结构树 PT；

输出：调序的源语言句子 RS；

步骤 1 层次遍历 PT，对遍历到的每一个节点 P，记录以 P 为根结点的孩子节点 C 的个数 n。

步骤 2 若 $n=2$，则记录该子树结构 $P \rightarrow C_1C_2$；若 $n>2$，则记录该子树结构下的所有伪子树结构 $P \rightarrow C_iC_{i+1}$，其中 $1 \leqslant i < n$。

步骤 3 将记录的子树结构与原规则进行匹配。若匹配成功，则按重排序规则进行左右子树结构的调整。当遍历完最后一个结点后，将该树的叶子节点按顺序输出就得到 RS。

在图 3 中可以看到，有两个子树结构符合原规则“VP→VV NP”，在应用了重排序规则“VP→NP VV”后，“谢谢 您 的 好 意”调整为“您 的 好 意 谢谢”，“叫 一 辆 出租汽车”调整为“一

辆 出租汽车 叫”，这分别与蒙古语“SAYIN SANAG_A-DV TANI BAYARLAL_A”和“TAKSI DAGVDAHV”的语序保持得更一致。

4 实验配置和结果分析

实验所用的语料库是由第五届全国机器翻译研讨会提供的67288个汉蒙句对作为训练集，400个汉语句子作为测试集，其中每个测试句子对应4个参考译文；实验使用小规模的源语言抽取词性标注序列模板，且限制模板中每个序列对应的频率数大于1；使用Stanford parser[13]工具对源语言进行依存分析和短语结构分析；使用SRILM[14]工具并采用改进的Kneser-key平滑算法进行3元语言模型的训练，使用GIZA++[15]工具并采用启发式方法进行词对齐的训练。

在使用评测工具进行评测之前，通过拉蒙转换算法将翻译的出的蒙古语句子从拉丁的形式转换为传统的形式。转换后的评分结果如表2所示，其中“基线”表示的是标准的基于短语的统计机器翻译系统，“自动抽取规则”表示的是使用由源语言短语结构树和词对齐限制自动抽取出的规则进行预处理的翻译系统，“序列预处理”表示的是经过词性标注序列模板进行预处理的翻译系统，“规则N预处理”表示的是使用与表1中相对应的重排序规则进行预处理的翻译系统。

表2 评分结果

类型	BLEU评分(%)	NIST评分
基线	22.74	5.5466
自动抽取规则	22.96	5.5432
序列预处理	23.59	5.6189
规则3预处理	23.06	5.6038
规则4预处理	23.84	5.6119
规则5预处理	23.90	5.6518
规则2预处理	24.24	5.6444
规则1预处理	24.45	5.6977

从表1可以看到，使用自动抽取的重排序规则进行预处理的翻译系统，其BLEU评分与基线相差不多，仅多了0.22个点，表明了该方法对源语言进行调序的效果不明显；基于词性标注序列模板进行预处理的系统，其BLEU评分比基线有很大提高，高出了0.85个点，表明了该方法是有效果的；基于手动书写重排序规则进行预处理的系统，其BLEU评分比基线有较为显著地提高，特别是使用规则5、规则2和规则1时提高的更显著，最多提高了1.71个点，表明了该方法是更有效果的，能更为显著地提高翻译系统的性能。

然而这两种调序方法都有一定的不足。对于第一种调序方法，能够弥补由依存分析错误或者依存重排序规则片面所导致相应语序没有调整的不足，但对译文质量的提高没有第二种调序方法显著；对于第二种调序方法，不仅可以实现短语和短语之间的语序调整，还可以实现单词和短语之间的语序调整，但是当使用不止一个重排序规则进行预处理所得到的评分结果并没有达到预想的全部提高，而第一种调序方法不会出现这种情况。另外，比如在图3中汉语“需要 很 长 的 时间”和“抚养 妻子 子女”也应该交换顺序的，但是这两种调序方法都捕捉不到这样的语序调整。

5 结论

本文提出了两种源语言重排序模型，一种是利用依存信息和词性标注信息形成词性标注序列模板，在模板的基础上来调整源语言，另一种是利用短语结构信息和词性标注信息形成重排序规则，在重排序规则和短语结构子树的基础上来调整源语言。从实验和翻译结果可以看到，这两种

方法都能提高译文的质量，特别是第二种方法提高的更显著。

然而这两种方法都有一定的局限性。第一种方法会出现由于有用的词性标注序列被过滤掉或者不全导致一些符合词性标注序列的源语言部分的语序没有得到相应调整的问题，第二种方法会出现由于规则中的短语结构标签部分发生规则嵌套导致一些句子进行过度调整的问题。下一步我们将利用序列的上下文信息和规则的上下文信息，并结合最大熵分类的方法来解决这些问题，更进一步地去提高统计机器翻译系统的性能和译文质量。

参 考 文 献

[1] Richard Zens, and Hermann Ney. A Comparative Study on Reordering Constraints in Statistical Machine Translation[C]. In proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 2003: 144-151.

[2] 薛永增, 李生, 赵铁军等. 短语统计机器翻译的句法调序模型[J]. 通信学报, 2008, 29(1): 7-14.

[3] 侯宏旭, 刘群, 李锦涛. 一种基于短语的汉蒙统计机器翻译与调序模型[J]. 高技术通讯, 2009, 19(5): 475-479.

[4] Fei Xia and Michael McCord. Improving a Statistical MT System with Automatically Learned Rewrite Patterns[C]. In Proceedings of the 20th International Conference on Computational Linguistics, 2004: 508-514.

[5] Michael Collins, Philipp Koehn, and Ivona Kucerova. Clause Restructuring for Statistical Machine Translation[C]. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 2005: 531-540.

[6] 张家俊, 宗成庆. 融入句型信息的汉英双向调序模型[C]. 宗成庆. 机器翻译研究进展——第四届全国机器翻译研讨会论文集. 中国科学院自动化研究所: 宗成庆, 2008: 222-230.

[7] Chao Wang, Michael Collins and Philipp Koehn. Chinese Syntactic Reordering for Statistical Machine Translation[C]. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007: 737-745.

[8] Nizar Habash. Syntactic Preprocessing for Statistical Machine Translation[C]. In Proceedings of the Machine Translation Summit XI, 2007: 215-222.

[9] Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. Using a Dependency Parser to Improve SMT for Subject-Object-Verb Languages[C]. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009: 245-253.

[10] 孙广泛, 宋金平, 肖健等. 句法调序的统计机器翻译方法研究[J]. 计算机工程与应用, 2009, 45(36): 142-144.

[11] Karthik Visweswariah, Jiri Navratil, and Jeffrey Sorensen. Syntax Based Reordering with Automatically Derived Rules for Improved Statistical Machine Translation[C]. In Proceedings of the 23rd International Conference on Computational Linguistics, 2010: 1119-1127.

[12] Young-Suk Lee, Bing Zhao, and Xiaoqiang Luo. Constituent Reordering and Syntax Models for English-to-Japanese Statistical Machine Translation[C]. In Proceedings of the 23rd International Conference on Computational Linguistics, 2010: 626-634.

[13] http://nlp.stanford.edu/software/lex-parser.shtml.

[14] Andreas Stolcke. SRILM-An Extensible Language Modeling Toolkit[C]. In Proceedings of the 7th International Conference on Spoken Language Processing, 2002: 901-904.

[15] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase-based Translation[C]. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003: 48-54.