基于分析和生成的复述与 SMT 语料扩展

和 为，刘 挺

（哈尔滨工业大学 计算机科学与技术学院，150001 哈尔滨）

摘 要：为了解决统计机器翻译语料对调序现象覆盖不足的问题，采用复述方法对语料进行扩展. 提出了一种基于依存分析和句子生成的复述方法. 对句子进行依存分析得到依存树，然后从依存树生成多个自然语言句子. 生成的句子与原句相比没有词汇上的改变，但可以在词序方面进行变换. 实验表明方法在不引入额外资源的前提下，有效缓解了语料覆盖不足的问题，提高了机器翻译质量.

关键词：复述；统计机器翻译；依存分析；句子生成

中图分类号：TP391.2 **文献标志码**：A **文章编号**：0367－6234(2013)05－0045－06

Parse-realize based paraphrasing and SMT corpus enriching

HE Wei, LIU Ting

(School of Computer Science and Technology, Harbin Institute of Technology, 150001 Harbin, China)

Abstract: To resolve the low-coverage problem of the statistic machine translation training corpus, a dependency parsing and sentence realization based paraphrasing method is proposed. The input sentence is first parsed into a dependency tree, and then the tree is realized into multiple natural language sentences. Although the generated sentences have the same lexical words, the expressions of word orders are re-arranged. The experiments shows that the paraphrasing method can be used to enlarge the bilingual corpus for statistic machine translation and the method efficiently relieves the low-coverage problem of training corpora without any extra resources, finally the translation quality is improved.

Key words: paraphrase; statistic machine translation; dependency parsing; sentence realization

复述(Paraphrase)，是对相同语义的不同表达[1]. 它是人类语言中固有的一种语言现象. 例如："我最喜欢玫瑰花"和"玫瑰花是我的最爱"就是一对互为复述的句子. 双语平行语料在统计机器翻译(SMT)中有至关重要的作用. 一般来说，双语平行语料的数量越多，机器翻译的质量就越好. 文献[2]指出双语语料的规模每增加 1 倍，SMT 的翻译得分(BLEU)就提高 2 点. 然而可获取的双语语料规模有限，尤其对于一些小语种如泰语、越南语等，获取双语资源更加困难. 双语语料的不足则会导致训练的统计机器翻译模型的覆盖率较低.

为了解决双语平行语料不足导致的 SMT 覆盖率较低的问题，一些研究者尝试使用复述来扩展双语平行语料. 文献[3]通过基于统计的复述生成模型，整合多种复述资源对双语平行语料进行复述扩展. 这个方法需要引入额外的复述资源，但对于某些语言来说，复述资源本身也是比较难获取的. 文献[4]和文献[5]使用基于规则的方法对双语平行语料进行复述扩展，这种方法需要人工编写复述规则，需要专业的语言学知识，导致方法很难扩展到其他语言.

本文提出了一种基于分析和生成的复述方法，首先对句子进行依存分析得到依存句法树，然后再通过句子生成方法从依存树生成句子. 生成的句子与原句保持相同的语义，只是词序方面发

收稿日期：2012－05－18.

基金项目：国家自然科学基金面上资助项目(61073126, 61133012)；国家高技术研究发展计划重大资助项目(2011AA01A207).

作者简介：和 为(1982—)，男，博士研究生；
刘 挺(1972—)，男，教授，博士生导师.

通信作者：刘 挺，tliu@ir.hit.edu.cn.

生了变化,所以两者是复述关系,属于语序变换的复述类别. 如果在句子生成过程中保留概率最高的前 N 个结果,就可以得到原句的多个复述句. 这种方法的优点是不需要任何复述资源. 用这种方法对 SMT 的双语平行语料的源语言部分进行复述扩展,得到的源语言复述句与对应的目标语言句子可以组成新的平行句对.

1 基于分析和生成的复述

对句子进行依存分析得到依存树,然后从依存树重新生成句子,取概率最高的前 N 个句子. 其中 N 取值越高,获得的扩展语料的规模越大,但语料质量也随之下降. 一般来说,N 的取值为 2 ~ 10 之间. 这 N 个句子对原始句子来说,即是单词相同但是顺序不同的复述句.

1.1 依存树结构

依存句法分析将句子由一个线性序列转化为一棵结构化的依存分析树,通过依存弧反映句子中词汇之间的依存关系,例如:“这是武汉航空首次购买波音客机”,依存分析的结果如图 1 所示.

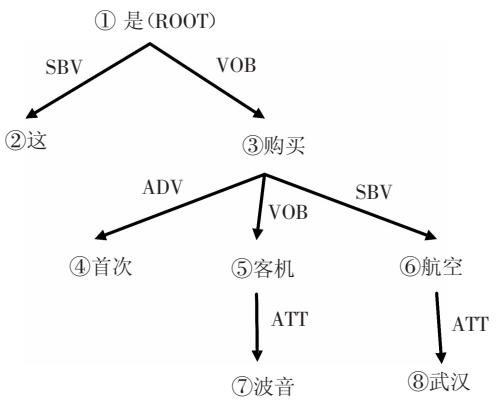

图 1 “这是武汉航空首次购买波音客机”的依存分析结果(依存树)

如图 1 所示,依存树上的每个节点表示 1 个单词. 单词“是”(节点①)是整棵依存树的根节点,其他节点(② ~⑧)都是根节点的子孙节点. 连接节点的依存弧表示单词之间的依存关系. 例如在图 1 中,从节点③ ~节点⑤ 有 1 条依存关系为 VOB 的依存弧,它表示单词“客机”依存于单词“购买”,并且它们之间是动宾类型的依存关系. 在依存树中,发出依存弧的节点被称为头节点,依存弧指向的节点称为依存节点.

1.2 句子生成

依存树只是反映了单词之间的依存关系,并没有记录单词之间的位置关系. 给定一棵依存句法树,可以通过句子生成模型[6]生成自然语言句子,取得分最高的 N 个结果,就得到了原句的多个复述句. 本文将从树(子树)结构生成单词序列的过程称为序列化.

1.2.1 分治策略

为了将依存树的树状结构转换成自然语言句子的线性结构,需要考虑全部节点所有的排列顺序然后选择概率最高的序列,这是一个 NP 问题. 为了使问题可求解,假设子树的序列化只与子树内部的节点有关,与其他子树无关. 基于这个假设,句子生成算法可以采用分治的策略将依存树切分成若干个深度为 1 的子树,并自底向上的对所有子树进行序列化操作. 以图 1 为例,首先将最底层的子树⑤ 和⑥ 序列化,然后处理子树③,最后进行子树① 的序列化.

1.2.2 相对位置假设

在依存结构中,依存关系的类型代表了节点之间的语法或语义关系. 例如,在动宾关系(VOB)中,头节点是一个动词,而依存节点是该动词的宾语;在定中关系(ATT)中,依存节点是头结点的修饰词. 在中文这种语序限定比较严格的语言中,一旦确定了依存关系的类型,头结点和依存节点的相对位置往往是一个固定的顺序. 例如在中文中,宾语总是出现在它的支配动词的后面;而修饰词总是出现在它修饰对象的前面. 本文给出一个相对位置假设:头结点和依存节点之间的相对位置仅由依存关系的类型决定.

为了验证上述假设,本文统计了哈工大依存树库(HIT – CDT)中所有依存弧的头结点和依存节点之间的相对位置. 为了描述方便,使用文献[7]中的定义,如果一个依存节点在句子中的位置在头结点之前,则称为前依存节点;如果依存节点的位置在头结点之后,则称为后依存节点. 表 1 给出了 HIT – CDT 树库(共 10 000 句)中每种依存关系的前依存节点和后依存节点的数量.

从表 1 可以看出,相对位置假设可以覆盖除 IS 类型(独立结构)之外的所有依存关系类型. 例如,100% 的 ATT 关系的依存节点都是前依存节点,99.9% 的 VOB 关系的依存节点都是后依存节点. 唯一的例外是 IS 类型(独立结构),它有 794 个(86.4%)前依存节点和 125 个(13.6%)后依存节点. 但是本文可以忽略这部分例外,将所有的 IS 类型的依存节点都按前依存节点处理. 这是因为:

1) IS 类型依存弧数量较少,在语料中仅占依

存弧总数的 0.5%，不满足相对位置假设的 IS 依存弧仅占总数的 0.07%；

2）IS 类型主要的标注对象的是位于括号内的说明性文本，这部分文本的位置本身就比较灵活，一般不影响生成句子的语义.

基于相对位置假设，可以根据依存关系将子树的叶节点分成前依存节点和后依存节点两类，在每一类之内进行序列化处理，这样在提高准确率的同时减少了计算开销.

表 1　HIT-CDT 树库前依存节点和后依存节点数量统计

依存关系	描述	前节点	后节点
ADV	状中结构	1	25 977
APP	同位关系	807	0
ATT	定中关系	0	47 040
CMP	动补结构	2 931	3
CNJ	关联结构	0	2 124
COO	并列关系	6 818	0
DC	依存分句	197	0
DE	“的”字结构	0	10 973
DEI	“得”字结构	131	3
DI	“地”字结构	0	400
IC	独立分句	3 230	0
IS	独立结构	125	794
LAD	前附加关系	0	2 644
MT	语态结构	3 203	0
POB	介宾关系	7 513	0
QUN	数量关系	0	6 092
RAD	后附加关系	1 332	1
SBV	主谓关系	6	16 016
SIM	比拟关系	0	44
VOB	动宾关系	23 487	21
VV	连动结构	6 570	2

1.2.3　句子生成模型

对子树的序列化，首先基于相对位置假设对依存节点进行分类，接下来通过对数线性模型寻找概率最高的前依存节点的序列和后依存节点的序列为

$$p_\lambda(r \mid t) = \frac{1}{Z_\lambda(t)}\exp\Big[\sum_{m=1}^{M}\lambda_m h_m(r,t)\Big].$$

这里 $Z_\lambda(t)$ 是一个归一化因子为

$$Z_\lambda(t) = \sum_{r' \in Y(t)}\exp\Big[\sum_{m=1}^{M}\lambda_m h_m(r',t)\Big].$$

使用的特征包括单词模型、依存关系模型和头节点模型，分别描述了单词序列、依存关系序列和头结点单词序列的概率. 具体计算与语言模型相似，不同的是分别对单词序列、依存关系序列和头结点单词序列建模并计算概率. 下面给出 3 个模型的计算公式.

1）依存关系模型. 设 $\bar{D}_1^I$ 为子树所有节点的一个序列化结果候选，依存关系模型计算 I 个节点的依存关系的 n-gram 概率（文中取 $n = 4$）. 设 $\bar{R}_k$ 为节点 $\bar{D}_k$ 的依存关系，则

$$P(\bar{R}_1^I) = P(\bar{R}_1 \cdots \bar{R}_I) = \prod_{k=1}^{I} P(\bar{R}_k \mid \bar{R}_{k-n+1}^{k-1}).$$

2）单词模型. 设 $\bar{S}_1^I$ 为子树序列化之后的子树单词串序列，设 $\bar{S}_1^I$ 由单词 $w_1, \cdots, w_m$ 组成. 则单词模型概率的计算公式为

$$P(\bar{S}_1^I) = P(w_1 \cdots w_m) = \prod_{k=1}^{m} P(w_k \mid w_{k-n+1}^{k-1}).$$

3）头结点模型. 头结点模型计算头结点单词的 n-gram 概率，设头结点单词序列为 $\bar{H}_1^I$，则计算公式为

$$P(\bar{H}_1^I) = P(H_1 \cdots H_I) = \prod_{k=1}^{I} P(H_k \mid H_{k-n+1}^{k-1}).$$

2　SMT 双语语料扩展

设双语平行语料为 $\boldsymbol{S} = \{(e_1, f_1), (e_2, f_2), \cdots, (e_n, f_n)\}$，其中：$e_1, \cdots, e_n$ 为源语言句子；$f_1, \cdots, f_n$ 为目标语言句子. 对句对 (e_i, f_i) 中的 e_i 进行复述，得到得分最高的 k 个复述句子，将这些句子按照得分降序排序，记为 $e_{i1}, e_{i2}, \cdots, e_{ik}$，这些句子与 f_i 组成新的双语句对.

2.1　扩展语料权重估计

双语句对 (e_i, f_i) 经过复述扩展之后，得到了 k 组新的句对，$\{(e_{i1}, f_i), (e_{i2}, f_i), \cdots, (e_{ik}, f_i)\}$. 对语料 $\boldsymbol{S}$ 中所有的句对都进行扩展后，按照排序，将新生成的句对划分成 k 个语料集合 $\boldsymbol{S}_1 \cdots \boldsymbol{S}_k$. 设 $\boldsymbol{S}_j = \{(e_{1j}, f_1), \cdots, (e_{nj}, f_n)\}$，复述扩展得到的语

料可以用矩阵表示为

$$\begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_2 \\ \vdots \\ \boldsymbol{S}_k \end{bmatrix} = \begin{bmatrix} (e_{11}, f_1) & (e_{21}, f_2) & \cdots & (e_{n1}, f_n) \\ (e_{12}, f_1) & (e_{22}, f_2) & \cdots & (e_{n2}, f_n) \\ \vdots & \vdots & & \vdots \\ (e_{1k}, f_1) & (e_{2k}, f_2) & \cdots & (e_{nk}, f_n) \end{bmatrix}.$$

对于语料集合 $\boldsymbol{S}_1, \cdots, \boldsymbol{S}_k$，根据它的排序来计算权重为

$$w(\boldsymbol{S}_j) = \frac{1}{k+j}, \quad (1 \leqslant j \leqslant k).$$

在不同的语料集合上分别训练翻译模型，然后将各个翻译模型中的短语对按照不同的权重进行合并. 通过对权重的调整实际上控制了不同质量的复述扩展语料对最终训练翻译模型的贡献.

2.2 SMT 输入的词网格扩展

基于依存分析和句子生成的复述方法不仅可以用来扩展 SMT 的训练语料，也可以在翻译过程中对 SMT 的输入句子进行扩展. 将一个句子扩展成多个复述句子，然后将多个句子合并，构建成词网格(Lattice)的形式，最后由 SMT 解码器进行翻译解码.

2.2.1 最小替换子串

设 SMT 的源语言输入句子为 s_0，生成的 m 个复述句子为 $s_1 \sim s_m$. 为了将 s_0 与 m 个复述句子合并构造成一个 Lattice，首先需要将复述句子转换成最小替换子串的形式. 最小替换子串是一个三元组：< MIN_RP_TEXT, COVER_START, COVER_LEN >. 其中：MIN_RP_TEXT 为完成复述变化需要替换的最短文本串；COVER_START 为替换的起始位置；COVER_LEN 为替换长度. 例如，源语言输入句子为 s_0 = “美丽 可爱 的 小 女孩”，复述句 s_1 = “可爱 美丽 的 小 女孩”，则最小替换子串为 < 可爱美丽, 0, 2 >，表示用本文串”可爱美丽”从位置 0 开始替换，替换的文本长度为 2 个词.

2.2.2 构建 Lattice

在把复述句转换成最小替换子串之后，可以使用短语复述的方法构造 Lattice[8]：给定输入句子(看做单词序列) $\{w_1, \cdots, w_N\}$ 作为输入，两个短语 $\alpha = \{\alpha_1, \cdots, \alpha_p\}$ 和 $\beta = \{\beta_1, \cdots, \beta_q\}$ 分别是 $P_1 = \{w_x, \cdots, w_y\}$ $(1 \leqslant x \leqslant y \leqslant N)$ 和 $P_2 = \{w_m, \cdots, w_n\}$ 的复述短语. 通过 2 个步骤可以构造 Lattice.

1)将原始句子构造成 Lattice. Lattice 中包含 $N+1$ 个节点 $(\theta_k, 0 \leqslant k \leqslant N)$ 和连接它们的 N 条边 $w_i (1 \leqslant i \leqslant N)$.

2) 为每个复述生成额外的节点和边. 以 α 为例，首先生成 $p-1$ 个节点，然后生成连接 θ_{x-1}，$p-1$ 个节点和 θ_{y-1} 的 p 条边，标记为 $\alpha_j (1 \leqslant j \leqslant p)$.

不断重复步骤 2)直到完成了为所有复述短语生成了新的 Lattice 节点，具体过程如图 2 所示.

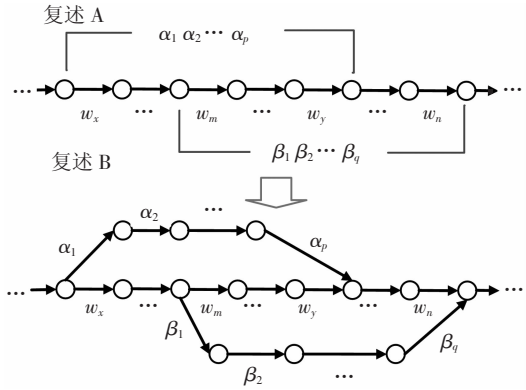

图 2 复述短语构建 Lattice 示例

3 实验结果与分析

3.1 实验数据

在实验中使用语言技术平台(LTP) http://ir.hit.edu.cn/ltp/对中文进行分词和依存分析. 在 SMT 方面，使用 Moses[9] 作为基线系统，使用 GIZA++[10] 进行词对齐，使用最小错误率训练算法[11]进行参数训练，语言模型的训练使用了 SRILM 工具 http://www.speech.sri.com/projects/srilm/(5 元语言模型)，评测方法使用 BLEU[12]. BLEU 是近年来使用最广泛的机器翻译自动评测标准，通过统计翻译结果中与参考译文匹配的 n-gram 数量，计算 n-gram 的加权平均准确率. 具体的计算公式为

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

式中：BP 为长度惩罚因子；p_n 为 n-gram 的准确率 $(1 \leqslant n \leqslant N$，实验中取 $N = 4)$；w_n 为 n-gram 的权重，一般按均匀分布取 $w_n = 1/N$. 在中 - 英方向上进行了实验，使用的训练集如表 2 所示.

表 2 SMT 训练语料

语料	句对数	中文单词数/k	英文单词数/k
BETC	19 972	174	190
PIVOT	20 000	162	196
HIT	80 868	788	850
CLDC	190 447	1 167	1 898
Tanaka	149 207	–	1 375

如表 2 所示，使用的双语语料包括 IWSLT 2008 的 BTEC 语料和 PIVOT 语料，哈工大双语语料 http://mitlab.hit.edu.cn/index.php/resources/

29-the-resource/111-share-bilingual-corpus. html. 和 CLDC 中英双语语料（CLDC-LAC-2003-004，CLDC-LAC-2003-006）. 语言模型的训练除了使用双语语料的英文部分之外，还使用了 Tanaka 语料.

使用的测试集语料如表 3 所示. 本文选择 CSTAR03-test 和 IWSLT06-dev 作为开发集，使用 IWSLT04-test、IWSLT05-test、IWSLT06-dev 和 IWSLT07-test 作为测试集.

表 3　中-英方向测试集语料

类型	语料	句子数	参考答案数
开发集	CSTAR03 test set	506	16
	IWSLT06 dev set	489	7
测试集	IWSLT04 test set	500	16
	IWSLT05 test set	506	16
	IWSLT06 test set	500	7
	IWSLT07 test set	489	6

3.2　实验结果

对中 - 英平行训练语料的中文句子首先进行依存分析得到依存句法树，然后使用句子生成算法从依存树生成复述句. 如果生成的复述句与原句完全相同，则去除这个结果. 最终取得分最高的前 3 个复述结果（$k = 3$）. 通过复述对中 - 英平行训练语料进行扩展. 设原始中 - 英平行训练语料为 $\boldsymbol{S}_0$，根据扩展语料权重估计描述的方法，分别用生成的复述句构造新的平行训练语料 $\boldsymbol{S}_1$，$\boldsymbol{S}_2$ 和 $\boldsymbol{S}_3$. 4 个语料分别训练翻译短语表（phrase table），然后根据不同的权重把 4 个短语表中的短语对（phrase pair）合并. 其中，$\boldsymbol{S}_0$ 的权重设为 1，其他 3 个语料的权重根据扩展语料权重估计的权重计算公式得到.

将使用 $\boldsymbol{S}_0$ 训练得到的模型作为基线模型（baseline），然后用该模型与复述扩展语料 $\boldsymbol{S}_1$，$\boldsymbol{S}_2$，$\boldsymbol{S}_3$ 训练的模型合并，得到了复述扩展模型. 在复述扩展模型的基础上，又尝试将 SMT 输入句子进行复述扩展并构造 Lattice 的结果. 实验结果如表 4 所示.

表 4　SMT 结果（BLEU）　%

模型	iwslt04	iwslt05	iwslt06	iwslt07
基线模型	53. 53	58. 87	27. 65	39. 77
复述扩展模型	54. 32	59. 40	28. 12	40. 95
复述扩展 + Lattice	54. 65	59. 91	28. 23	41. 30

3.3　分析与讨论

从实验结果可以看出，通过复述对 SMT 的双语训练语料进行扩展，扩展后的 SMT 模型翻译质量比基线系统提高 0. 47% ~1. 18%. 在 SMT 的输入句子也进行复述扩展并构建 Lattice 之后，翻译质量进一步提升，BLEU 提高了 0. 58% ~1. 53%.

表 5 对比了复述扩展模型和基线模型的短语对数量，并以 iwslt07 测试集为例比较了两个模型的 n-gram 覆盖率.

表 5　翻译模型对 iwslt07 的覆盖率对比

模型	短语对总数	iwslt07 覆盖率/%			
		1-gram	2-gram	3-gram	4-gram
基线模型	1 552 941	98. 7	75. 6	40. 8	19. 7
述扩展模型	1 969 439	99. 0	78. 2	42. 7	20. 8

如表 5 所示，经过复述扩展的翻译模型中的短语对总数比基线模型增加了 27%. 扩展后的模型对测试集的 1 ~4 元的覆盖率均有提高. 本文方法中，覆盖率的增长都是在不引入任何额外资源的情况下获得的. 结合表 4 的数据可以看出，本文提出的方法有效的提高了翻译模型的覆盖率，最终导致了翻译质量的提高. 这种方法不借助额外资源，可以很容易的用于一些小语种的双语语料扩展.

在表 5 中 iwslt07 的 1-gram 覆盖率也有微弱的提升（0. 3%）. 理论上说，通过分析和生成产生的复述句子不会生成新的一元单词，但实际上词序的调整造成了双语语料的词对齐结果发生变化，最终导致有一些新的一元单词翻译对被挖掘出来.

4　结　论

1）提出了一种基于依存分析和句子生成的复述方法，可以对句子的词序进行调整，生成多个复述句，使语言现象变丰富.

2）通过基于依存分析和句子生成的复述方法，对统计机器翻译的双语平行语料进行扩展，增强了翻译模型的覆盖率，最终提高了翻译质量.

参考文献

[1] BARZILAY R, MCKEOWN K R. Extracting paraphrases from a parallel corpus[C]//Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2001: 50 - 57.

[2] KOEHN P, OCH F J, MARCU D. Statistical phrase-based translation [C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language. Stroudsburg, PA: Association for

Computational Linguistics, 2003: 48 – 54.

[3] HE Wei, ZHAO Shiqi, WANG Haifeng, *et al.* Enriching SMT training data via paraphrasing [C]// Proceedings of the 5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand: IJCNLP, 2011: 803 – 810.

[4] BOND F, NICHOLS E, APPLING D S, *et al.* Improving statistical machine translation by paraphrasing the training data [C]//Proceedings of the International Workshop on Spoken Language Translation (IWSLT). USA: Hawail, 2008:150 – 157.

[5] NAKOV P. Improved statistical machine translation using monolingual paraphrases[C]//Proceedings of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial Intelligence. The Netherlands: IOS Press Amsterdam, 2008: 338 – 342.

[6] HE Wei, WANG Haifeng, GUO Yuqing, *et al.* Dependency based Chinese sentence realization [C]// Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg, PA: Association for Computational Linguistics, 2009:809 – 816.

[7] COVINGTON M A. A fundamental algorithm for dependency parsing [C]//Proceedings of the 39th Annual ACM Southeast Conference. New York: ACM, 2001: 95 – 102.

[8] DU Jinhua, JIANG Jie, WAY A. Facilitating translation using source language paraphrase lattices [C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2010: 420 – 429.

[9] KOEN P, HOANG Hieu, BIRCH A, *et al.* Moses: open source toolkit for statistical machine translation [C]// Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. ACL Demo and Poster Sessions. Stroudsburg, PA: Association for Computational Linguistics, 2007: 177 – 180.

[10] OCH F J, NEY H. Improved statistical alignment models[C]//Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2000: 440 – 447.

[11] OCH F J. Minimum error rate training in statistical machine translation [C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2003: 160 – 167.

[12] PAPINENI K, ROUKOS S, TWARD T, *et al.* BLEU: a method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 311 – 318.

（编辑 张 红）