北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.032

使用源语言复述知识改善统计机器翻译性能

苏晨 张玉洁† 郭振 徐金安

北京交通大学计算机学院，北京 100044; † 通信作者, E-mail: yjzhang@bjtu.edu.cn

摘要 为了缓解双语语料不足所带来的翻译知识欠缺问题，提出基于复述技术的翻译框架。此框架利用第 3 种语言获取带有概率的复述知识表，以 Lattice 表示输入句子的多种复述形式，扩展解码器使之可以对 Lattice 形式的输入进行解码，将复述知识作为特征加入到对数线性模型的目标函数。在保持原始翻译知识表不变的情况下，此框架不仅可以增大短语翻译表对源语言现象的覆盖率，也能够增加候选译文的表现形式的多样性。在 3 个不同规模训练集上的对比实验结果表明，在训练语料规模最小的情况下(10 K[①])，系统性能有明显提升(BLEU+1.4%)；在训练语料规模最大的情况下(1 M)，系统性能也取得了性能上的提升(BLEU+0.32%)。

关键词 复述知识；短语翻译表；特征；解码器

中图分类号 TP391

Improved Statistical Machine Translation with Source Language Paraphrase

SU Chen, ZHANG Yujie†, GUO Zhen, XU Jin'an

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044; †Corresponding author, E-mail: yjzhang@bjtu.edu.cn

Abstract The performance of statistical machine translation (SMT) suffers from the insufficiency of parallel corpus. To solve the problem, the authors propose a paraphrase based SMT framework with three solutions: 1) acquiring paraphrase knowledge based on a third language; 2) expressing multiple paraphrases of input sentence in a lattice and modifying decoder to be able to process it; 3) integrating paraphrase knowledge as features into log-linear model. In this way, not only more expressions in source language can be covered, but also more expressions in target language can be generated as candidate translations. To verify proposed method, experiments are conducted on three training data sets with different sizes, and evaluate the improvement of the performance of SMT system contributed by paraphrasing. Experimental results show that the translation performance is improved significantly (BLEU+1.4%) when the parallel corpus is small (10 K), and a good performance (BLEU+0.32%) is also achieved when parallel corpus is large enough (1 M).

Key words paraphrase; phrase translation table; features; decoder

在统计机器翻译(SMT)中，系统性能往往受限于平行语料规模的大小。对于训练语料中未出现的词汇 (OOV)，SMT 系统的通常做法是将其保留在翻译结果中，严重影响了译文质量。同时，人工构建大规模高质量平行语料费时费力，而自动构建平行语料又难以保证其质量。为了解决这一问题，研究人员开展了利用复述技术的机器翻译方法的研究，近年来成为研究的热点之一[1–3]。

复述是在同种语言内，表达与原始形式语义相同，内容不同的表现形式。机器翻译系统无法翻译的

国家国际科技合作专项(2014DFA11350)、国家自然科学基金(61370130)和北京交通大学人才基金(2011RC034)资助
收稿日期: 2014–06–30; 修回日期: 2014–10–29; 网络出版时间: 2014-12-01 09:26

本文提及的语料规模(10 K, 100 K, 1 M)单位是句对，如: 10K 表示 10000 句对的语料。

句子，可以通过复述处理得到不同的表现形式，如果翻译系统可以翻译其中的某种表现形式，那么输入句子就可以获取译文。复述技术可以在一定程度上改善由于翻译知识不足导致的无法翻译问题。复述知识可以从第三语种的平行语料或单语语料中获取，相比扩展训练数据的平行语料，这些语言资源的获取更加容易。

本文以英中翻译为例，提出基于复述技术的翻译框架，主要研究利用日语作为中间语言获取英语复述表的方法以及利用复述特征的解码算法。本文以 NTCIR[①]英中翻译任务为例，在 3 个不同规模的训练集上设计对比实验，分析短语翻译表的规模由小变大的过程中，复述处理对系统性能提升的贡献程度。

1 使用复述知识的翻译框架

对于统计机器翻译而言，短语翻译表是主要的翻译知识。但是由于平行语料规模的限制，所获取的短语翻译表很难覆盖所有的测试用例，对译文质量的影响主要表现以下两方面。

1) OOV: 当测试语料中存在未知词汇时，SMT 系统通常不做任何处理，因此测试语料中未知词汇会影响译文质量。

2) 义项不全: 短语翻译表难以覆盖某一词汇的所有语义的翻译知识，这会导致测试语料中的句子不能被正确地翻译。例如对英文句子“Now let me talk about Article II on Labor Law.”使用翻译系统进行翻译时，虽然短语表中“Article”有多条译文选项，但由于缺少“Article→条款”，译文也很难令人满意。

针对这两个问题，本文提出基于复述技术的 SMT 框架，如图 1 所示，其中虚线框中的部分是本文的主要工作。相比于传统的 SMT 框架，本文框架增加了复述生成模块，对解码器进行了扩展。对于源语言句子，首先利用复述短语表生成复述 Lattice，然后作为解码器的输入进行解码。

复述生成模块对源语言句子的任意长度的字符串都将查询复述短语表，生成由源语言句子和相应复述构成的格图 Lattice，如图 2 所示。其中，带有标号的节点表示词语的分界，从节点 i 指向节点 $i+1$ ($i\geqslant 0$)的实线表示序号为 $i+1$ 的单词，它的信息包括原始短语和权重(复述概率)；节点 i 指向节点 $i+k$($k\geqslant 1$)的虚线表示从单词 $i+1$ 到单词 $i+k$ 组成的短语的复述，它的信息包括复述短语和权重(复述概率)。复述 Lattice 保存了输入句子的多种复述形式，有助于解码阶段得到丰富的候选译文；解码算法依据复述的权重对复述的译文重新评分。

在构建 Lattice 过程中，权重的设置至关重要。通过分析发现: 1) 权重惩罚过大会导致由复述知识获取的译文得分较低，难以被 SMT 系统选中，在翻译知识缺乏时无法显著改善翻译性能；2) 权重惩罚过小则导致由复述知识产生的噪声影响变大，尤其对于翻译知识比较充足的 SMT 系统，复述知识反而会降低其翻译性能。为了解决这个问题，本文提出将复述知识作为新的特征，加入到对数线性模型中，通过在开发集参数训练，使复述知识的权重自动适应 SMT 系统。

与本文所采用的复述知识的翻译框架相比，文献[2]只是将源语言句子中的 OOV 替换为它的复述。尽管能够改善翻译系统的性能，但是它只解决

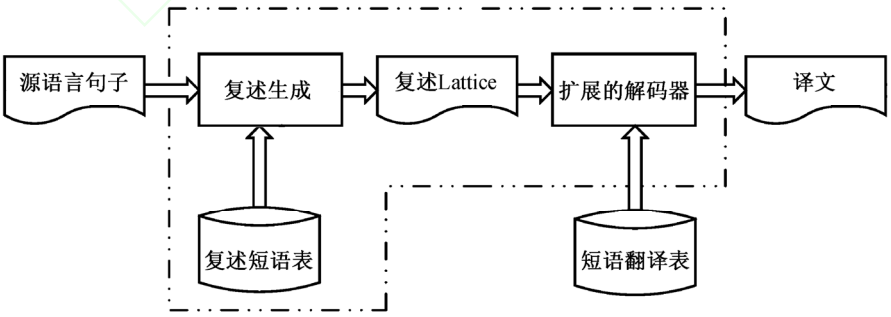

图 1　基于复述的翻译框架

Fig. 1　Framework of paraphrase-based SMT

①本文实验采用的数据来自 NTCIR 英中机器翻译评测数据(http://ntcir.nii.ac.jp/about/)。

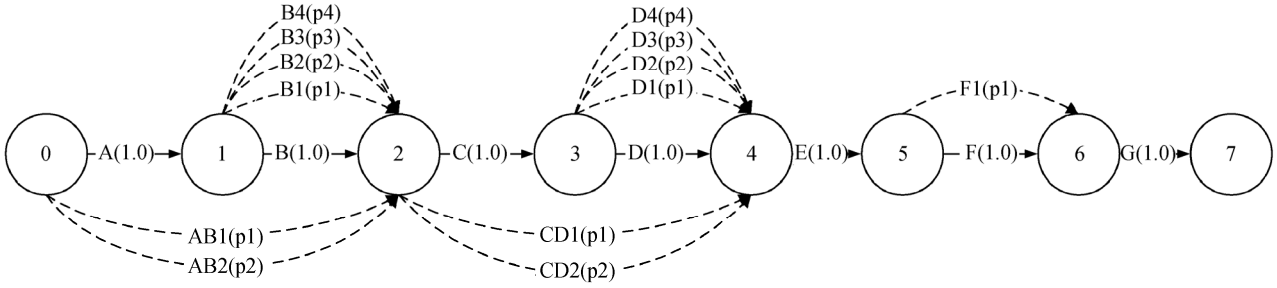

图 2　输入句子的复述 Lattice

Fig. 2　Using Lattice graph to denote input sentence's different paraphrases

上述问题 1, 没有涉及问题 2。针对问题 2, 文献[1]使用复述 Lattice 进行解码, 但是所用的复述权重是固定的, 无法实现自适应。

2　复述短语表的获取

2.1　复述短语表的获取方法

本文以 NTCIR 英中翻译任务为例, 研究基于复述的统计机器翻译。复述知识的获取方法主要分为从单语语料获取的方法和从双语语料获取的方法[4–5]。在我们的任务中, 尽管作为训练语料的 NTCIR 英中双语语料数量有限(1 M), 但是相同领域上其他语言与英语的双语语料很丰富, 比如 NTCIR 英日双语语料有 3 M 的规模。本文采用利用双语语料获取复述知识的方法, 从 NTCIR 英日平行语料获取英语复述知识。

在英日平行语料中, 我们通过日语作为桥梁获得英语短语之间的复述关系。如果不同的英文短语 e_2 和 e_1 都翻译成相同的日语短语 jp, 那么英文短语 e_2 与 e_1 互为复述。复述概率可通过式(1)得到。

$$\text{para}(e_2|e_1) \approx \sum_{jp} p(e_2|\text{jp}) \cdot p(\text{jp}|e_1) \tag{1}$$

其中, $p(\text{jp}|e_1)$表示英语短语 e_1 翻译为日语短语 jp 的概率, $p(e_2|\text{jp})$是日语短语 jp 翻译为英文短语 e_2 的概率。短语的翻译概率可以使用极大似然估计得到。

$$p(e_2|\text{jp}) \approx \frac{\text{count}(e_2, \text{jp})}{\sum_{e_2} \text{count}(e_2, \text{jp})} \tag{2}$$

$$p(\text{jp}|e_1) \approx \frac{\text{count}(\text{jp}, e_1)}{\sum_{jp} \text{count}(\text{jp}, e_1)} \tag{3}$$

count(e, jp)表示在平行语料中, 英文短语 e 和日文短语 jp 对齐的次数。

复述权重的另外一种计算方法可以采用式(4)[1]:

$$\text{Weight}(e) = \frac{1}{k+i} (1 \leqslant i \leqslant k) \tag{4}$$

其中 k 设定为 7, i 是当前复述 e 依据 para 的排名。

2.2　复述短语表对短语翻译表的扩展

本文提出的基于复述的翻译框架, 可以解决上述问题 1 和 2。对于问题 1, 将短语翻译表无法翻译的短语复述成另外一种形式, 获得译文, 提高短语翻译表对语言现象的覆盖率。对于问题 2, 通过丰富输入短语的表现形式, 增加候选译文多样性,提高短语翻译表对于正确译文的覆盖率。下面通过评价在这两方面覆盖率的提升, 展示复述短语表对短语翻译表的扩展效果。

本文使用 NTCIR 英中平行语料, 以 10 K,100 K 和 1 M 规模的数据作为训练语料获取短语翻译表, 然后统计它们对 NTCIR 测试语料(2 K)的覆盖率, 评测结果列于表 1 中箭头(→)左侧。对比覆盖率变化发现: 当训练数据规模较小时, 增加训练语料的规模能有效提升短语翻译表的覆盖率, 如 10 K→100 K, 1 元短语的覆盖率提升 13.88%(从 77.19 增加到 91.07%); 而当短语翻译表的规模达到一定程度时, 训练语料规模的增加对于翻译知识的覆盖率提升不明显, 如 100 K→1 M, 语料规模扩大了 10 倍, 而 1 元短语的覆盖只增加了 4.34% (从 91.07 增加到 95.41%)。随后统计加入复述知识后, 短语翻译表对测试语料的覆盖率, 结果列于表中箭头(→)右侧。通过对比发现, 在训练语料较小的情况下, 加入复述知识对于提升翻译知识的覆盖率有极大帮助, 如在 10 K 平行语料上构建的短语表, 在加入复述知识后其覆盖率提升 12.78% (77.19%→89.97%); 当训练语料规模较大时, 提升效果不明显, 如在 1 M 平行语料上构建的翻译表中加入复述知识, 覆盖率只提升了 0.14% (95.41%→95.55%)。

表 1　不同规模训练语料上构建的短语翻译表对于测试语料的覆盖率以及加入复述知识后短语翻译表的覆盖率

Table 1　Test data's coverage from phrase translation table, which are based on training data with different scales, and in the case of introducing the paraphrasing knowledge

测试数据 *N*-元短语(数量)	翻译表的覆盖率(原始短语翻译表→经复述扩张后的翻译表)/%		
	10 K 训练语料	100 K 训练语料	1 M 训练语料
1 元(6274)	77.19→89.97	91.07→93.27	95.41→95.55
2 元(28993)	35.57→67.52	62.95→74.08	80.15→82.59
3 元(42937)	12.97→37.11	30.01→42.63	48.91→53.09
4 元(46974)	4.20→14.94	11.50→18.39	22.77→25.91
5 元(47316)	1.50→5.55	4.620→7.36	10.66→11.98
6 元(46389)	0.55→2.12	1.99→3.09	5.36→5.99
7 元(44918)	0.24→0.81	0.97→1.37	3.08→3.31

根据第 2 节第 2 个问题可知, 一些短语尽管在短语表中存在, 但义项不全, 找不到合适的译文, 也会造成翻译质量下降。引入复述知识后, 这个问题在一定程度上会得到缓解。为了探究复述知识对于该问题的改善程度, 即理想译文与参考译文的相似度变化, 本文设计了另外一个实验。理想译文是候选译文中与参考译文相似度最高的译文。本文采用的相似度指译文与参考译文的最长公共子序列长度与参考译文长度的比例, 以汉字为单位。为了选取理想译文, 使用 CKY 算法模拟解码过程, 以相似度作为目标函数, 根据式(5)和(6)计算理想译文与参考译文的相似度。式(5)中 $f(e_i^j, c_l^m)$ 表示英文短语 e_i^j 的理想译文与参考译文 c_l^m 最长公共子序列的长度, 其中 i, j, $k(i<k<j)$是英文句子中单词之间的分界点编号, e_i^j 表示英文句子中从第 i 个分界点到第 j 个分界点之间的单词组成的短语, 同理 l, m, $n(l<n<m)$是参考译文 c 的汉字分界点; $\mathrm{Length}(c_l^m)$ 是短语 c_l^m 的汉字数目。式(6)中 S 表示测试语料中句子的数目, $f(e_s, c_s)$表示语料中第 s 个英文句子的理想译文与它的参考译文的最长公共子序列的长度。本实验模拟解码过程, 找到测试语料的理想译文, 并计算它与参考译文的相似度, 结果见表 2。

$$f(e_i^j, c_l^m) = \max \begin{cases} \mathrm{Length}(c_l^m), & \text{如果 } e_i^j \to c_l^m \text{ 在翻译知识中,} \\ f(e_i^k, c_l^n) + f(e_k^j, c_n^m), & \text{单调调序,} \\ f(e_i^k, c_n^m) + f(e_k^j, c_l^n), & \text{交换调序,} \\ f(e_{i+1}^j, c_l^m), & \text{如果 } e_i^{i+1} \to \mathrm{null} \text{ 在翻译知识中,} \\ f(e_i^{j-1}, c_l^m), & \text{如果 } e_{j-1}^j \to \mathrm{null} \text{ 在翻译知识中,} \\ f(e_i^j, c_{l+1}^m), & \text{如果 } c_l^{l+1} \to \mathrm{null} \text{ 在翻译知识中,} \\ f(e_i^j, c_l^{m-1}), & \text{如果 } c_{m-1}^m \to \mathrm{null} \text{ 在翻译知识中,} \end{cases} \tag{5}$$

$$\mathrm{Similarity}(c, c_{\mathrm{ref}}) = \sum_{s=1}^{S} f(e_s, c_s) \Big/ \sum_{s=1}^{S} \mathrm{Length}(c_s) \tag{6}$$

由表 2 可知, 通过复述知识增加了原始短语翻译表的义项, 在一定程度上解决了在短语翻译表中找不到合适译文的问题。尤其在翻译知识比较匮乏时, 如在 10 K 平行语料上构建短语翻译表, 通过加入复述知识, 理想译文与参考译文的相似度上升 9.64% (82.82%→92.46%), 随着翻译知识逐渐变得丰富, 复述知识对译文的改善变弱, 在 1 M 平行语料上构建的翻译知识在加入复述知识后, 理想译文与参考译文相似度提高 2.72% (94.31%→97.03%)。由此可见, 即使对于本实验的 1 M 训练语料的机器翻译系统, 通过引入复述知识, 译文质量仍有 2.72%的提升空间。

表 2　不同规模训练语料构建的短语翻译知识与加入复述知识的情况下, 理想译文与参考译文的相似度

Table 2　Similarities of ideal translation and reference translation, in the cases of phrase translation knowledge built by training data with different scales and with additional paraphrase knowledge

翻译知识来源	相似度/%		
	10 K 训练语料	100 K 训练语料	1 M 训练语料
短语翻译表	82.82	91.22	94.31
短语翻译表+复述知识	92.46	95.88	97.03

3 引入复述特征的解码算法

基于短语的统计机器翻译系统采用对数线性模型进行解码，如式(7)：

$$\hat{c}=\arg\max_{c}\{\Pr(c|e_1)\}=\arg\max_{c}\{\sum_{m=1}^{M}\lambda_m h_m(c,e_1)\} \quad (7)$$

h_m 表示不同特征的目标函数。在基于短语的机器翻译系统中，有 4 个与短语翻译有关的特征：正向短语翻译 $h_{\text{Tran}}(c, e_1)$、反向短语翻译 $h_{\text{VerTran}}(c, e_1)$、正向词汇化 $h_{\text{Lex}}(c, e_1)$和反向词汇化 $h_{\text{VerLex}}(c, e_1)$。

在解码阶段加入源语言复述知识 $e_1\rightarrow e_2$，首先将源语言短语 e_1 复述成 e_2，然后再使用 e_2 查询短语翻译表。本文将复述知识加入到目标函数后，得到新的目标函数如式(8)~(11)所示。

$$\hat{h}_{\text{Tran}}(c,e_1)=\log\hat{p}(c|e_1)=\log\left[\text{para}(e_2|e_1)^{\alpha_1}\cdot p(c|e_2)\right] \quad (8)$$

$$\hat{h}_{\text{VerTran}}(c,e_1)=\log\hat{p}(e_1|c)=\log\left[\text{para}(e_1|e_2)^{\alpha_2}\cdot p(e_2|c)\right] \quad (9)$$

$$\begin{aligned}\hat{h}_{\text{Lex}}(c,e_1)&=\log\hat{\text{Lex}}(c|e_1)\\&=\log\left[\text{para}(e_2|e_1)^{\alpha_3}\cdot\text{Lex}(c|e_2)\right]\end{aligned} \quad (10)$$

$$\begin{aligned}\hat{h}_{\text{VerLex}}(c,e_1)&=\log\hat{\text{Lex}}(e_1|c)\\&=\log\left[\text{para}(e_1|e_2)^{\alpha_4}\cdot\text{Lex}(e_2|c)\right]\end{aligned} \quad (11)$$

其中 para($e_2|e_1$)表示 e_1 复述成 e_2 的概率，依据式(1)计算得到。式(8)和(10)合并得到式(12)；式(9)和(11)合并得到式(13)。

$$\begin{aligned}&\lambda_{\text{Tran}}\cdot\hat{h}_{\text{Tran}}(c,e_1)+\lambda_{\text{Lex}}\cdot\hat{h}_{\text{Lex}}(c,e_1)\\&=\lambda_{\text{Tran}}\log[\text{para}(e_2|e_1)^{\alpha_1}\cdot p(c|e_2)]+\\&\quad\lambda_{\text{Lex}}\log[\text{para}(e_2|e_1)^{\alpha_3}\cdot\text{Lex}(c|e_2)]\\&=\lambda_{\text{Tran}}\log p(c|e_2)+\lambda_{\text{Lex}}\log\text{Lex}(c|e_2)+\\&\quad(\lambda_{\text{Tran}}\cdot\alpha_1+\lambda_{\text{Lex}}\cdot\alpha_3)\log\text{para}(e_2|e_1)\\&=\lambda_{\text{Tran}}\cdot h_{\text{Tran}}(c,e_2)+\lambda_{\text{Lex}}\cdot h_{\text{Lex}}(c,e_2)+\\&\quad\lambda_{\text{Para}}\cdot h_{\text{Para}}(e_1,e_2)\end{aligned} \quad (12)$$

$$\begin{aligned}&\lambda_{\text{VerTran}}\cdot\hat{h}_{\text{VerTran}}(c,e_1)+\lambda_{\text{VerLex}}\cdot\hat{h}_{\text{VerLex}}(c,e_1)\\&=\lambda_{\text{VerTran}}\log[\text{para}(e_1|e_2)^{\alpha_2}\cdot p(e_2|c)]+\\&\quad\lambda_{\text{VerLex}}\log[\text{para}(e_1|e_2)^{\alpha_4}\cdot\text{Lex}(e_2|c)]\\&=\lambda_{\text{VerTran}}\log p(e_2|c)+\lambda_{\text{VerLex}}\log\text{Lex}(e_2|c)+\\&\quad(\lambda_{\text{VerTran}}\cdot\alpha_2+\lambda_{\text{VerLex}}\cdot\alpha_4)\log\text{para}(e_1|e_2)\\&=\lambda_{\text{VerTran}}\cdot h_{\text{VerTran}}(c,e_2)+\lambda_{\text{VerLex}}\cdot h_{\text{VerLex}}(c,e_2)+\\&\quad\lambda_{\text{VerPara}}\cdot h_{\text{VerPara}}(e_1,e_2)。\end{aligned} \quad (13)$$

与传统模型相比，基于复述的模型实际上引入了两个新特征：正向复述特征 $h_{\text{Para}}(e_1, e_2)$和逆向复述特征 $h_{\text{VerPara}}(e_1,e_2)$。$\alpha=\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$是不定参数，$\lambda_{\text{Para}}=(\lambda_{\text{Tran}}\cdot\alpha_1+\lambda_{\text{Lex}}\cdot\alpha_3)$ 与其他参数独立，同理 $\lambda_{\text{VerPara}}=(\lambda_{\text{VerTran}}\cdot\alpha_2+\lambda_{\text{VerLex}}\cdot\alpha_4)$也是独立的。由此可见，复述知识作为两个新的特征加入了对数线性模型。最优的特征权重依据式(14)可以使用最小错误率训练[6]得到。

$$\hat{\lambda}_1^M=\arg\max_{\lambda}\{\sum_{s=1}^{S}\log\Pr(c,e_1)\} \quad (14)$$

分析可知，当 e_1 与 e_2 相同时，para($e_1|e_2$)= para($e_2|e_1$)=1.0，此时 $h_{\text{Para}}(e_1, e_2)=h_{\text{VarPara}}(e_1, e_2)=0$，而且与复述短语 e_2 相关的特征全部变为与原短语 e_1 相关的特征，例如 $h_{\text{Tran}}(c|e_2)\rightarrow h_{\text{Tran}}(c|e_1)$，此时加入复述特征的模型退化为传统模型。当 $\lambda_{\text{Para}}=1.0$ 且 $\lambda_{\text{VerPara}}=0.0$ 时，该系统与文献[1]的系统极为相似。

4 评测实验与结果分析

4.1 实验数据

为了对比系统的有效性，本文设计了 3 个不同翻译系统进行对比。第一个系统是传统的短语模型[7]，记作 Baseline；第二个系统是依照文献[1]的方法实现的系统，记作 Du System；最后一个系统是将复述知识作为新特征加入到 SMT 的系统，记作 Our System。本文使用 NiuTrans1.3.0①搭建短语模型 SMT 系统。单词对齐的结果由 GIZA++②训练得到，然后使用 grow-diag-and-final③启发式算法进行对称化；短语模型的最大长度设置为 7；对数线性模型参数训练方法使用最小错误率。

实验在英中翻译系统上进行验证，为了详细比较不同规模数据上系统性能的差异，本文在 3 个不同规模的训练集开发机器翻译系统：10 K，100 K 和 1 M 英中训练语料。1 M 规模的训练语料是 NTCIR 中英训练语料；100 K 和 10 K 语料分别在 1 M 的训练语料中随机获取得到；由于 1 M 的训练语料的内容分布不均匀，本文采用伪随机获取策略：将语料平均分成很多组，每组中句子编号是连续的，每组

① http://www.nlplab.com/NiuPlan/NiuTrans.ch.html
② https://code.google.com/p/giza-pp/
③ http://www.statmt.org/moses/?n=FactoredTraining.AlignWords

随机取一个句子组成新语料。

NTCIR 的中英开发集和测试集语料各 2000 句对，每个句子只有一个参考译文，所有翻译的系统共用这套开发集和训练集。译文的评测指标采用 BLEU 值[8]。

本文使用式(1)获取复述知识，采用 NTCIR 英日平行语料，规模是 3 M。首先，在训练语料上获得相应的英日短语翻译表，然后依据短语翻译表中的 $p(c|e_1)$和 $p(e_2|c)$，计算复述知识“$e_1 \rightarrow e_2$”正向复述概率 para($e_2|e_1$)和反向复述复述概率 para($e_1|e_2$)。在获取过程加入剪枝策略以过滤可信度较低或者不会被测试语料使用的复述知识，剪枝策略主要包括以下 4 方面：1) 如果在开发集和测试集中均未出现英文短语 e_1，则去掉复述短语知识 $p(e_2|e_1)$；2) 如果短语翻译表中未出现英文短语 e_2，则去掉复述短语知识 $p(e_2|e_1)$；3) 式(1)中的 $p(c|e_1)$和 $p(e_2|c)$必须高于阈值 0.01；4) 对于每一个 e_1 只保留得分最高的 50 条复述知识。

4.2 实验结果与分析

本文在不同规模的数据集上进行 3 组实验，对比 Baseline 系统、Du System 和 Our System 的性能。评测结果列于表 3。通过观察发现在 10 K 规模的训练数据集上，加入复述知识后，SMT 系统性能得到较明显的提升。相对于 Baseline 系统，Du System 的 BLEU 值提升 1.03%，Our System 系统提升 1.4%。当训练数据规模扩大到 100 K 时，Du System 和 Our System 的性能相对于 Baseline 系统分别提高了 0.29%和 0.03%。当训练数据规模增加至 1 M，Du System 相对于 Baseline 系统 BLEU 值明显降低(–0.73%)，而 Our System 系统 BLEU 值提升 0.32%，相对于 Du System 提升 1.05%。结果表明 Our System 的方法不仅能在训练语料最少的情况下，改善翻译系统性能，并且在训练语料充足的 SMT 系统上表现了较好的性能。

表 3 不同规模训练语料上搭建的 SMT 系统性能比较：Baseline 系统、Du System 和 our System

Table 3 Comparison for the performance of SMT systems, which are based on training data with different scales, including Baseline system, Du system and our system

统计机器翻译系统	BLEU 值(增减量) /%		
	10 K 训练数据	100 K 训练数据	1 M 训练数据
Baseline	35.69	40.39	44.16
Du System	36.72(+1.03)	40.68(+0.29)	43.43(–0.73)
Our System	37.09(+1.40)	40.42(+0.03)	44.48(+0.32)

对不同翻译系统的译文进行详细对比和分析发现，当翻译知识比较匮乏时，复述知识对翻译知识的扩充效果很明显，因而有利用改善 SMT 系统的性能。以 10 K 规模数据集开发的 3 个系统的 1-best 译文为例，当翻译源语言句子“Particularly, screw bases of the type e14, e26 or e27 are frequently used for lamps.”时，“Particularly”的参考译文是“尤其是”，而短语翻译表中并未出现“Particularly→尤其是”，Baseline 系统将其翻译为“具体地说”，而通过复述关系可以得到翻译知识：“Particularly→In particular→尤其是”。Du System 和 Our System 通过复述知识都得到了正确的译文。随着训练语料规模的增加，短语翻译表变得更加完善。通过 2.2 的统计结果可知，复述知识对短语翻译表的拓展效果变弱，而由复述知识带来的噪声对 SMT 翻译性能的影响愈发明显。以 1 M 训练数据上 3 个不同翻译系统的 1-best 翻译结果为例，当翻译英文句子“The lighting device 1 further has an optical fiber 5 which is coupled to the solid-state light source 4.”时，在参考译文中，“which is coupled”被翻译为“它被耦合”，但是在 Du System 中，将复述知识“which is coupled→coupled”和翻译知识“coupled→耦合”结合，得到的译文相比于“which is coupled→它被耦合”得分更高，因此 Du System 采用“耦合”作为“which is coupled”的译文。在短语翻译表的规模较大时，Our System 对复述特征惩罚力度较大，因而“which is coupled”被正确翻译。

通过对比实验和结果分析，本文提出的将复述知识作为特征的方法不仅在训练语料较少时能够提升系统性能，而且在训练语料充足的情况下，避免了由复述知识的噪声引起的 SMT 系统性能下降的问题。

5 结语

针对训练语料有限，翻译知识不足的问题，本文提出了基于复述技术的翻译框架，主要解决了 3 个问题：1) 将目光转向英语和其他语种的丰富的平行语料，利用第 3 种语言获取带有概率的英语复述知识；2) 以 Lattice 形式保存输入句子的所有的复述表现形式，扩展解码器使之可以对 Lattice 形式

的输入进行解码; 3) 将复述知识作为特征, 加入到对数线性模型的目标函数。在保持原始翻译知识表不变的情况下, 这种框架提高了短语翻译表对源语言现象的覆盖率, 也增加了更多的候选译文以接近参考译文。本文在 3 个不同规模训练集上设计对比实验, 分析翻译知识表由小到大过程中, 复述技术对系统性能的贡献程度。实验结果证明, 无论短语翻译知识匮乏或丰富, 本文的方法都取得了不错的效果。

本文评测中所采用的测试集只有一个参考译文, 在具有更多参考译文的测试集上评测分析, 或者增加 Meteor 和 TER 进行评测, 将能更全面评价本文的方法。本文报告了在英中翻译上我们方法的有效性, 验证该方法在其他语言对翻译上的效果将是今后的工作。此外, 本文只针对源语言端进行了复述, 目标端的复述处理也可以改善 SMT 的性能, 将其作为特征加入对数线性模型可以作为下一步的研究工作。

参考文献

[1] Du Jinhua, Jiang Jie, Way A. Facilitating translation using source language paraphrase lattices // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Massachusetts: Association for Computational Linguistics, 2010: 420–429

[2] Callison-Burch C, Koehn P, Osborne M. Improved statistical machine translation using paraphrases // Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. New York: Association for Computational Linguistics, 2006: 17–24

[3] 赵世奇, 刘挺, 李生. 复述技术研究. 软件学报, 2009, 20(8): 2124–2137

[4] Madnani N, Dorr B J. Generating phrasal and sentential paraphrases: a survey of data-driven methods. Computational Linguistics, 2010, 36(3): 341–387

[5] Wu Hua, Zhou Ming. Synonymous collocation extraction using translation information // Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Sapporo: Association for Computational Linguistics, 2003: 120–127

[6] Och F J. Minimum error rate training in statistical machine translation // Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Sapporo: Association for Computational Linguistics, 2003: 160–167

[7] Koehn P, Och F J, Marcu D. Statistical phrase-based translation // Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Atlanta: Association for Computational Linguistics, 2003: 48–54

Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th annual meeting on association for computational linguistics. Philadelphia: Association for Computational Linguistics, 2002: 311–318