基于 Transfer 和 Triangulation 融合的中介语统计机器翻译方法 *

王强[1]，杜权[1]，肖桐[1]，朱靖波[1]
（1.东北大学自然语言处理实验室，辽宁省 沈阳市 110819）

摘要：为了解决在构建统计机器翻译系统过程中所面临的双语平行数据缺乏的问题，本文提出了一种新的基于中介语的翻译方法，称为 transfer-triangulation 方法。本文方法可以在基于中介语的翻译过程中，结合传统的 transfer 方法和 triangulation 方法的优点，利用解码中介语短语的方法改进短语表。本文方法是在使用英语作为中介语的德-汉翻译任务中进行评价的。实验结果表明，相比于传统的基于中介语方法的基线系统，本文方法显著提高了翻译性能。

关键词：统计机器翻译; 基于中介语的统计机器翻译；中介语；质量控制因子
中图分类号：TP391 **文献标识码：**A

Transfer-Triangulation Method for Pivot-based Statistical Machine Translation

Qiang Wang[1], Quan Du[1], Tong Xiao[1], Jingbo Zhu[1]
(1.Northeastern University NLP Lab, Shenyang, Liaoning 110819, China)

Abstract: This paper presented a new approach to pivot-based translation between a language pair with poor bilingual data, referred to as transfer-triangulation method, which takes the best of both typical transfer method and triangulation method for pivot-based translation and decodes pivot phrases to improve phrase table. Our approach was evaluated on German-Chinese translation task with English as the pivot language. Experimental results show that our method achieves significant improvement over baseline pivot-based methods.

Key words: statistical machine translation; pivot-based statistical machine translation; pivot language; quality control factor

1 引言

构建性能优异的统计机器翻译系统通常需要数百万乃至更多的双语平行数据进行训练。然而在实际应用时，除少量数据资源丰富的语言对外（如英汉、英阿），大多数语言对往往面临双语平行数据资源缺乏的问题（如德汉）。

为此，研究人员提出了基于中介语的统计机器翻译，其核心思想是：通过与源语和目标语均存在大规模平行语料的第三方语言，间接地满足源语-目标语的平行数据的要求。两个典型的基于中介语的统计机器翻译方法分别为 Transfer 方法[1]和 Triangulation[2,3]方法。Transfer 方法是句子级的中介语方法，核心思想是先将源语句子翻译为中介语句子，再将中介语句子翻译为目标语句子。其缺点是翻译过程需要解码两次，不但更耗时且存在翻译错误蔓延的问题。而 Triangulation 方法是短语级的中介语方法，核心思想是分别训练源语-中介语短语翻译表T_{s-p}、中介语-目标语的短语翻译表T_{p-t}，再利用相同的中介语短语进行

*收稿日期： 定稿日期：
基金项目：国家自然科学基金青年基金项目 (No.61300097);国家自然科学基金面上项目(No. 61272376);国家自然科学基金重点项目(No. 61432013)
作者简介：王强（1990—），男，硕士，主要研究方向为机器翻译；杜权（1989—），男，博士研究生，主要研究方向为机器翻译；肖桐（1982—），男，副教授，博士，主要研究方向为机器翻译；朱靖波（1973—），男，教授，博士生导师，主要研究方向为机器翻译。

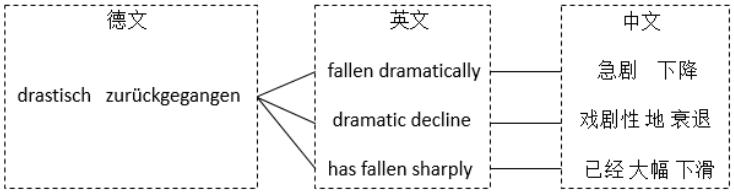


图 1 使用 Triangulation 方法进行源语-目标语翻译规则推导

短语表融合，构建出源语-目标语的短语表T_{s-t}。Triangulation 方法能够利用推导出的源语-目标语短语表直接进行翻译，避免了 Transfer 方法的解码两次的不足,并且其处理对象是短语，相比于句子有更大的灵活性，成为了目前中介语统计机器翻译的研究热点。然而，在 Triangulation 方法中，只考虑了在T_{s-p}和T_{p-t}中共现的中介语短语，忽略了非共现的中介语短语（本文称这种类型的中介语短语为中介语断点）。这将导致产生大量的互译性不高的噪声翻译规则，干扰解码器的译文选择过程，并且还存在源语短语丢失的问题。

针对上述问题，本文提出一种基于 Transfer 和 Triangulation 融合的中介语方法，其核心思想是利用短语级而不是句子级的 Transfer 方法，将原本被忽略的中介语断点翻译成目标语，形成中介语-目标语的翻译规则，从而将中介语断点转化成非断点。本文提出的方法能够利用传统 Triangulation 方法中忽略的中介语断点信息改善推导出的短语表,从而提高整体翻译性能。在以英文作为中介语的德-汉翻译任务中，本文的方法相比于传统的 Transfer 方法和 Triangulation 方法，BLEU-4 分别提高 4.74 和 0.84。

2 基于中介语的统计机器翻译

2．1 Transfer 方法

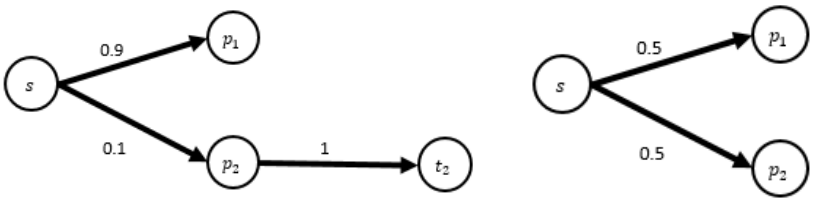
Transfer 方法是一种句子级的中介语方法。首先分别利用源语-中介语、中介语-目标语双语平行数据训练出源语-中介语翻译系统$S_{\mathrm{s-p}}$以及中介语-目标语的翻译系统S_{p-t}。给定源语句子s，当进行源语-目标语的翻译任务时，利用$S_{\mathrm{s-p}}$先将s翻译成$m(m \geq 1)$个中介语结果，记作p_1，$p_2 \ldots p_m$，再通过S_{p-t}将每一个中介语结果$p_i(1 \leq i \leq m)$翻译为$n(n \geq 1)$个目标语译文，记作t_{i1}，$t_{i2} \ldots t_{in}$，共产生$m * n$个翻译结果，最后从中选择 *1best* 作为最终的翻译结果。由于 Transfer 方法中需要解码两次$(s \rightarrow p$和$p \rightarrow t)$，所以整体的解码时间更耗时，更关键的是，连续的解码将造成翻译错误的蔓延。

2．2 Triangulation 方法

Triangulation 方法是短语级的中介语方法。首先分别训练源语-中介语短语翻译表T_{s-p}、中介语-目标语的短语翻译表T_{p-t}，再利用相同的中介语短语进行短语表融合，构建出源语-目标语的短语表T_{s-t}，该过程如图 1 所示。在德-英短语表中，存在翻译规则 drastisch zurückgegangen → fallen dramatically，同时在英-汉短语表中，存在翻译规则 fallen dramatically → 急剧 下降,通过共现的英文短语fallen dramatically，能够推导出德-汉翻译规则 drastisch zurückgegangen → 急剧下降。同理，还可以推导出 drastisch zurückgegangen → 戏剧性 地 衰退、drastisch zurückgegangen → 已经 大幅 下滑。Triangulation 方法中最关键的问题是：如何给推导出的短语翻译规则进行特征打分，主要包括双向的短语翻译概率、双向的词汇化权重。给定源语短语s，目标语短语t，则在 Triangulation 方法中[2]，使用公式（1）对基于中介语的短语翻译概率∅进行建模：

$$\emptyset(\mathrm{s}|\mathrm{t}) = \sum_{\mathrm{p}} \emptyset(\mathrm{s}|\mathrm{p})\emptyset(\mathrm{p}|\mathrm{t}) \qquad (1)$$

使用公式（2）对基于中介语的源语-目标语的词对齐推导：



(a) 噪声翻译规则 (b) 源语短语丢失

图 2. 传统 Triangulation 方法存在的问题。直线上的数值表示短语翻译概率

$$A_{s-t} = \{(s,t)|\exists p: (s,p) \in A_{s-p} \& (p,t) \in A_{p-t}\} \quad (2)$$

其中，A_{s-p}、A_{p-t}、A_{s-t}分别表示源语-中介语、中介语-目标语、源语-目标语之间的词对齐信息。使用公式（3）计算词汇化权重进行[4]:

$$p_\omega(s|t,a) = \prod_{i=1}^{n} \frac{1}{\{j|(i,j) \in a\}} \sum_{\forall(i,\ j)\in a} \omega(s_i|t_j) \quad (3)$$

其中，

$$w(s|t) = \frac{count(s,t)}{\sum_{t'} count(s,t')} \quad (4)$$

在基于中介语的统计机器翻译中，可以使用公式（5）[2]对源语词汇和目标语词汇共现次数进行建模。

$$count(s,t) = \sum_{k=1}^{K} \emptyset_k(s|t) \sum_{i=1}^{n} \delta(s,s_i)\delta(t,t_{a_i}) \quad (5)$$

其中，K表示被推导出的规则总数；当$x = y$时，$\delta(x,y) = 1$，否则$\delta(x,y) = 0$。

使用上述的公式对推导出的短语规则进行特征打分后，便得到了完整的源语-目标语的短语翻译表。然后按照标准的基于短语的统计机器翻译方法，直接进行源语到目标语的翻译。

虽然 Triangulation 方法能够直接把源语翻译为目标语，避免了 Transfer 方法中多次解码造成的翻译错误蔓延问题。但是，该方法也面临其他的问题：

1) 产生互异性不高的噪声翻译规则。如图 2(a)所示，源语短语s翻译为中介语p_1的概率为 0.9，表示为$\emptyset(p_1|s) = 0.9$，同时$\emptyset(p_2|s) = 0.1$。在应用 Triangulation 方法时，由于高翻译概率的p_1无法翻译为任何目标语短语，则s只能通过低翻译概率的p_2推导，从而形成互译性不高的翻译规则$s \to t_2$，而这些噪声翻译规则将干扰解码器的译文选择过程。
2) 源语短语丢失。如图 2(b)所示，源语短语s对应的全部中介语短语p_1和p_2都无法翻译成任何目标语短语，导致s无法推导出目标语翻译规则，所以在应用 Triangulation 方法时，s将在被构建的源语-目标语短语表中丢失。

本文定义图 2(a)中的p_1，以及图 2(b)中的p_1和p_2为中介语断点，称这种现象为中介语断路。以上两个问题产生的主要原因都是由于中介语断路，所以本文的出发点就是通过解码中介语断点的方式将其转化成非断点，利用更多的中介语信息改善短语翻译表质量。

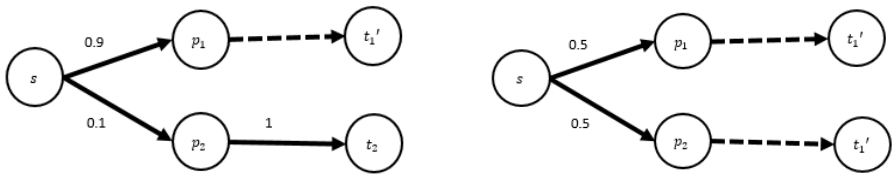
3 Transfer-Triangulation 方法

3．1 中介语断点

对于任意源语短语s，本文定义满足下列条件的中介语短语为中介语断点条件：

$$\exists p \in \bar{P}_s: < s,p > \in T_{s-p} \ \& \ \nexists t \in \bar{t}: < p,t > \in T_{p-t}$$

其中，T_{s-p}、T_{p-t}分别表示由源语-中介语、中介语-目标语平行数据训练得到的短语表，记T_{s-p}中的中介语短语集合为$\bar{P}_s$，T_{p-t}中的目标语短语集合为$\bar{t}$，$< x,y >$表示一条短语翻译规则$x \to y$。



(a) 缓解噪声翻译规则 (b) 缓解源语短语丢失

图 3. 利用解码中介语短语缓解上述问题示意图

在本文实验数据中，我们发现约 75%的中介语短语是断点。这说明大量的中介语短语无法在 Triangulation 方法中用来推导源语-目标语短语规则，造成可用规则的丢失，同时影响已推导出的翻译规则的概率估计。而出现中介语断路的原因是由于源语到中介语的双语训练语料和中介语到目标语的双语训练语料中，不可避免地存在如领域、语言习惯、表达方式等各种差异，最终反应到训练得到的短语翻译表中。所以，可以说中介语断路现象难以避免，而且会随着语料的相关性差异增大而越来越严重，而中介语断路现象本身也将影响 Triangulation 方法的性能。

基于上述分析，本文提出一种基于 Transfer 和 Triangulation 方法融合的中介语方法，如图 3 所示，其核心思想是利用短语级而不是句子级的 Transfer 方法，通过翻译（用虚线表示）中介语断点（图 3(a)中的p_1，图 3(b)中的p_1、p_2）得到目标语译文（图 3(a)中的t_1'，图 3(b)中的t_1'，t_2'，这里"'"表示该目标语短语是通过翻译断点得到的，而不是原本就存在于短语表中），形成中介语-目标语的翻译规则（如图 3(a)中的$s \to t_1'$，图 3(b)中的$s \to t_1'$、$s \to t_2'$），从而将中介语断点转化成非断点。本文提出的方法能够利用传统 Triangulation 方法中丢失的中介语断点信息改善推导出的短语表，从而提高整体翻译性能。

3．2 解码中介语断点

在本文中，如果解码中介语断点的结果出现未登录词（OOV），则定义解码失败。如果解码成功，则可以获得$n - best$个目标语译文t'_1，$t'_2 \ldots t'_n$，本文为简化计算，这里只取$n = 1$。因此，每解码成功一个中介语断点，即可产生一条新的中介语-目标语的翻译规则$p \to t'$。为了在 Triangulation 方法中应用新产生的$p \to t'$翻译规则，根据公式(1)-(5)可知，需要获得$p \to t'$的短语翻译概率$\phi(t'|p)$和词对齐信息$A_{p-t'}$。所以，解码中介语断点主要需要解决两个问题：

1） 如何计算$p \to t'$的短语翻译概率和词对齐

2）应该解码哪些中介语断点

本小节主要解决的是问题(1)。给定 D 是将中介语断点p翻译为目标语t'的完整推导过程，则使用公式(6)计算短语规则$p \to t'$的短语翻译概率$\phi(t'|p)$：

$$\phi(t'|p) = \prod_{d \in D} \phi(t'_d|p_d) \qquad (6)$$

$p \to t'$的词对齐推导算法描述如图 4 所示。算法核心思想是根据翻译推导的过程，依次将与推导对应的$span[i,j](j > i \geq 0)$、$span[j+1,k](k > j+1)$拼接，根据目标语拼接方向（正向或反向），更新$span[i,k]$的词对齐信息。图 4 中的 Step2 就是更新两个 span 词对齐的过程，Step3 是进行 span 拼接，得到更新词对齐后的更大的 span，并继续利用翻译推导信息继续更新词对齐。图 5 展示了使用本算法更新两个 span 词对齐结果的示例。

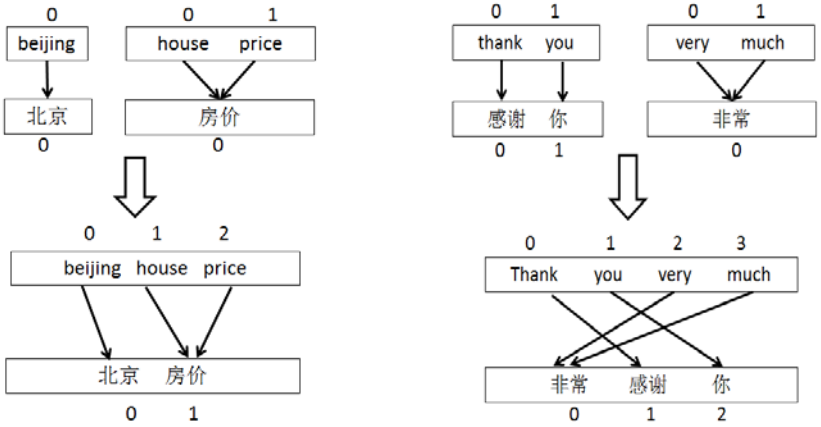
3．3 质量控制因子

本小节描述的是如何解决判断哪些中介语断点应该被解码的问题。直觉上，并不是所有中介语断点都对完善短语表有帮助。我们期望捕获的是在不可靠的短语推导过程中，没有被

输入： 中介语短语$P = p_{d1} \dots p_{dn}$，目标语短语$T = t_{d1} \dots t_{dn}$，P 与 T 之间的翻译推导$D = d_1 \dots d_n$，词对齐信息$\mathcal{A} = \mathcal{A}_{d1} \dots \mathcal{A}_{dn}, n \geq 2$
输出： 中介语短语 P 和目标语短语 T 之间的词对齐$\mathcal{A}_{P-T}$

Step1: 将两个起始推导d_1、d_2对应的 span 分别记为 L，R；翻译推导索引 i=2
Step2: if　L 与 R 是正向拼接
　　更新 R 中每一个源语对齐节点位置：加上 L 中源语片段的词数
　　更新 R 中每一个目标语对齐节点位置：加上 L 中目标语片段的词数
　if　L 与 R 是反向拼接
　　更新 R 中每一个源语对齐节点的位置：加上 L 中源语片段的词数
　　更新 L 中每一个目标语对齐节点的位置：加上 R 中目标语片段的词数
Step3: 将 L 和 R 拼接组成新的 span，作为 L
　翻译推导索引 i+1
　if $i \leq n$
　　将d_i对应的 span 作为 R，重复 Step2
　else
　　返回 L 的词对齐信息

图 4. 解码结果的词对齐推导算法



(a) Span 正向拼接　　(b) Span 反向拼接

图 5. 解码结果的词对齐推导算法示例

利用的高质量中介语断点。

该任务的进一步描述是：令函数$trans(x, y)$返回在短语表x中，y对应的所有翻译结果按正向短语翻译概率由大到小排好序的集合，则当给定源语短语s时，设$\bar{p} = \text{trans}(T_{s-p}, s)$，表示s对应的所有中介语翻译结果有序集合；设$\bar{t} = \bigcup_{p \in \bar{p}} \text{trans}(T_{p-t}, p)$，表示s对应的所有目标语翻译结果集合。其中，$\bar{p}$是由两部分组成：中介语断点有序集合$\bar{P}_b$及非断点有序集合$\bar{P}_{nb}$。而我们的目标是判断短语推导过程$I(s, \bar{p}, \bar{t})$是否可靠，若不可靠，则从$\bar{P}_b$中挑选子集$\bar{P}_{b'}$作为高质量中介语断点。其中涉及到如何衡量短语推导过程的可信度的问题，以及如何确定$\bar{P}_{b'}$的问题。

本文引入质量控制因子ψ的概念，利用推导产生的包含最大正向短语翻译概率的翻译规则所使用的中介语信息，衡量在给定源语短语s的推导过程的质量，其定义如下：

$$\psi(s, \bar{p}, \bar{t}) = \frac{\text{sort}(\arg\max_{p} \phi(t|s))}{\|\bar{p}\|} \qquad (7)$$

其中，$t \in \bar{t}$，$p \in \bar{p}$，$max(\phi(t|s))$表示的是由s推导出的短语规则集合中正向短语翻译概率最大的规则$s \rightarrow t_{max}$，则$arg\ \underset{p}{max}\ \phi(t|s)$表示 $s \rightarrow t_{max}$所经过的中介语短语集合$\bar{p}_{max}$，函数$\text{sort}(\bar{x})$表示将$\bar{x}$的元素按照短语翻译概率$\phi(x|s)$由大到小排序，返回排在第1位的中介语短语在$\bar{p}$中的位置，$\|\bar{p}\|$则表示$\bar{p}$中包含的中介语短语数。也就是说，质量控制因子实际是

表 1. 双语训练数据/开发集/测试集 数据说明

翻译方向	训练集	开发集		测试集	
	句对数	句对数	BLEU	句对数	BLEU
德-英	7,730,687	2000	22.65	2000	22.99
英-汉	9,806,071	1859	25.72	995	27.05

表 2. 系统训练得到的短语翻译表及 Triangulation 方法推导出的短语翻译表，其中 M 表示百万

翻译方向	Top30 短语翻译规则数	德语短语数	英文短语数
德-英	100M	33.9M	19.15M（包含）
英-汉	120M	N/A	39.69M（包含）
德-汉	812M	25.06M	4.95M（使用）

根据$s \to t_{max}$对应中介语短语集合$\bar{\mathrm{p}}_{\max}$衡量推导质量。如果$\bar{\mathrm{p}}_{\max}$中包含正向短语翻译概率越高的中介语短语,则$\mathrm{sort}(\bar{\mathrm{p}}_{\max})$越小，质量控制因子也越小，从而表示经过$\bar{\mathrm{p}}_{\max}$推导的翻译规则越可信。

利用质量控制因子计算出的推导过程可靠性，本文将所有推导分为如下 3 类：

1）丢弃型 *Discard*：$\forall \mathrm{p} \in \bar{\mathrm{p}}, \mathrm{t} \in \bar{\mathrm{t}}: < p \to t > \notin \mathrm{T}_{\mathrm{p-t}}$

2）低可信型 *Low*：$\exists \mathrm{p} \in \bar{\mathrm{p}}, \mathrm{t} \in \bar{\mathrm{t}}: < p \to t > \in \mathrm{T}_{\mathrm{p-t}}\ \&\ \psi(\mathrm{s}, \bar{\mathrm{p}}, \bar{\mathrm{t}}) > \lambda$

3）高可信型 *High*: $\exists \mathrm{p} \in \bar{\mathrm{p}}, \mathrm{t} \in \bar{\mathrm{t}}: < p \to t > \in \mathrm{T}_{\mathrm{p-t}}\ \&\ \psi(\mathrm{s}, \bar{\mathrm{p}}, \bar{\mathrm{t}}) \le \lambda$

其中λ是一个常数，表示质量控制因子的阈值,本文将高质量中介语断点集合$\bar{\mathrm{P}}_{b'}$表示为：

$$\bar{\mathrm{P}}_{b'} = \left\{ p \left| \frac{sort(p)}{\|\bar{\mathrm{p}}\|} \le \lambda, \mathrm{p} \in \bar{p} \right. \right\} \tag{8}$$

所以，本文解码中介语断点的定位是：通过解码$Discard$型推导中的中介语断点缓解源语短语 OOV 问题，通过解码Low型推导中低于λ的高质量中介语断点产生更多优质的翻译规则。

4 实验结果与分析

4．1 实验设置

德-英、英-汉系统使用的数据如表 1 所示。我们采用基于短语模型的 NiuTrans 开源工具[5]完成以英语为中介语的德-汉翻译任务。使用 GIZA++[6]工具获得双向词对齐结果，再使用“$grow-diag-final-and$”方法[4]进行词对齐对称化。抽取德-英短语对的长度设置为 3-3，英-汉短语对的长度设置为 3-5，则最终被推导出的德-汉短语长度为 3-5。对所有抽取的短语翻译表进行取 Top-N 处理，这里设置 N=30，即每一个源语短语对应的翻译候选最多为 30 个。分别使用 66,522,497 句和 42,946,518 句单语句子训练 5 元英文和中文的语言模型，均使用修正的 Kneser-Ney 平滑方法[7]。需要注意的是，在解码中介语断点时，使用的仍是上述语言模型，并没有因为解码结果是短语而做针对性优化。所有的特征采用最小错误率训练 MERT[8]进行参数调优。使用基于词的 BLEU-4[9]评价最终的翻译性能。

4．2 实验结果及分析

实验一 中介语断点比例

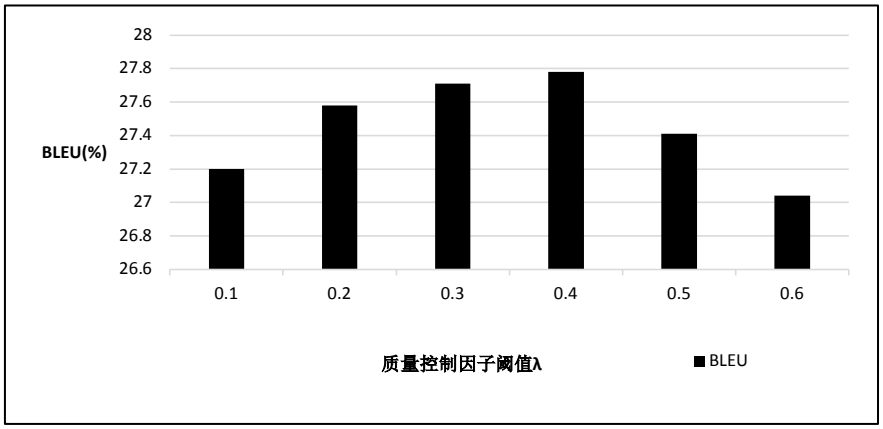


图 6. 质量控制因子阈值λ在开发集上对翻译结果的影响

表 3. Baseline 及处理不同推导类型在测试集上的翻译性能，*表示显著高于 Baseline

方法	BLEU	OOV 句子数	OOV 次数
Transfer	23.70	731	1029
Triangulation	27.60	731	1029
+ Discard	27.65	723	1009
+ Low ($\lambda = 0.4$)	28.44*	731	1029
+Discard &Low($\lambda = 0.4$)	28.36*	723	1009

应用本文的实验数据及设置，得到德-英、英-汉短语翻译表信息如表 2 第一行及第二行所示。然后使用 Triangulation 方法得到被推导出的德-汉短语翻译表，其信息如表 2 第三行所示。可以看到，德-英短语表中包含 1915 万唯一的英文短语；英-汉短语表中包含 3969 万唯一的英文短语,但只有 495 万条英文短语在 Triangulation 方法中被使用。

这里，我们以德-英的英文短语条数为参考，则断路的英文短语数为 1915-495=1420 万条，比例达（1420 万/1915 万）*100%=74.15%。也就是说，在德-英短语表中，有 74.15%的英文短语存在断路情况,这是一个在基于 Triangulation 的中介语统计机器翻译中普遍存在的问题。而本文的出发点正是想缓解中介语短语断路的问题。

实验二 质量控制因子阈值λ对翻译系能的影响

由于不同的质量控制因子阈值的设置，对判断需要解码的中介语短语数目有关，从而对改善短语翻译表产生影响，这里我们做了下列实验：在开发集上，通过改变阈值λ的取值，观察其对翻译结果的影响。实验结果如图 6 所示。

之所以呈现先增后减的趋势，本文分析结果是：如果质量控制因子阈值设置的过小，只有较少的高质量断路中介语短语被重新解码利用起来，对整体的翻译性能帮助并不明显。但如果设置质量控制因子的阈值过大，将会引入一些低质量的中介语短语，从而对翻译性能造成损害。这里我们看到$\lambda = 0.4$时翻译性能达到最高，所以后续的实验默认设置$\lambda = 0.4$。

实验三 不同推导类型对翻译性能的影响

我们对比了传统的 Transfer 方法、Triangulation 方法和本文提出的 Transfer-Triangulation 方法中处理不同推导类型的翻译性能结果，如表 3 所示。

由第一行和第二行可以看到，Triangulation 方法比 Transfer 方法翻译性能更好，BLEU

值上升了 3.9 个点。对于+*Discard*方法，在 2000 句的测试集上仅仅减少了8个未登录词，并没有如设想地缓解了未登录词问题。分析其中原因发现，对于大多数的包含源语未登录词的源语-中介语短语，相应的中介语短语也包含未登录词，从而导致解码的结果中也包含未登录词，造成解码失败。对于+*Low*方法，由于传统的 Triangulation 方法丢失了一些高质量的中介语短语，而本文方法能够有效利用这部分高质量的中介语短语进行短语翻译表改善，最终实验结果也证实了该想法的有效性。对于+*Discard&Low*方法同只+Low 的方法在 BLEU 上没有太多差异，且对 OOV 现象缓解的作用很小。

5. 相关工作

基于中介语的统计机器翻译的典型方法有两种：Transfer 方法和 Triangulation 方法。

对于 Transfer 方法，由 2.1 节分析可知，给定一个源语句子，最终会产生m * n个目标语的翻译结果。González-Rubio 和 Duh 等人[10,11]提出使用基于最小贝叶斯风险的系统融合方法去选择最优的翻译结果。

对于 Triangulation 方法，Kholy 等人[12]提出从词对齐信息中抽取两个与语言独立的特征，该特征指示了被推导出的源语-目标语翻译规则的可靠性。Tofigh 等人[13]提出利用基于中介语的上下文向量，从而计算被推导出的源语-目标语翻译规则间的短语相似度，并依据该相似度进行短语表过滤，从而起到过滤噪音规则的目的。朱晓宁等人[14]提出使用随机漫步方法获取潜在的源语-目标语短语路径，从而缓解源语未登录词问题。而后，朱晓宁等人[15]又提出在融合短语表前直接对源语-目标语的短语对共现次数进行估计的方法，避免了在短语推导时由于中介语断点导致破坏短语翻译概率空间的问题。Miura 等人[16]提出在进行短语规则推导时记录所使用的中介语信息，在进行源语-目标语的翻译过程中额外考虑中介语的语言模型特征。

另外，Michael 等人[17]探索了不同中介语的选择对系统的影响，英文更适合作为印欧语系及部分亚洲语言（如泰语、越语）之间的中介语。

不同于上述方法，本文提出的 Transfer-Triangulation 方法是将把 Transfer 方法应用于短语级，利用解码中介语短语的方法改善被推导出的短语表。

6 总结

本文提出 Transfer 和 Triangulation 融合的中介语统计机器翻译方法，通过应用短语级的 Transfer 方法，将高质量的中介语断点解码成相应的目标语短语，从而将中介语断点转换为非断点，使得 Triangulation 方法能够利用更多中介语信息，达到改善短语表、提高翻译性能的目的。其中，本文解决了计算解码结果短语翻译概率和词对齐问题，并提出了质量控制因子的概念，将使用 Triangulation 方法推导过程分为三类：丢弃型、低可信、高可信，利用质量控制因子阈值挑选 Triangulation 方法中无法使用的高质量中介语短语信息。实验结果表明，中介语短语断路现象是在应用 Triangulation 方法时普遍存在的问题，本实验中断路的中介语短语比例达 74.15%；随着质量控制因子阈值λ增大，翻译性能呈先上升后下降的趋势，原因在于：如果λ过小，只有较少的高质量断路中介语短语被解码，而如果λ过大，将会引入低质量的断路中介语短语，损害翻译性能；对低可信推导中的高质量中介语断点重新解码产生的推导规则，能够有效改善传统 Triangulation 方法推导出的短语表，减少了噪音翻译规则并且扩大了短语表的覆盖度，BLEU 值提高了 0.86 个点。但是对丢弃型推导的重解码处理并没有如预期有效缓解 OOV 问题，其原因在于源语如果包含 OOV，其相应的中介语短语也往往包含 OOV，从而造成解码失败。未来我们将探索如何将中介语的解码结果作为翻译特征帮助解码器选择正确的翻译选项。

参考文献

[1] Masao Utiyama and Hitoshi Isahara. 2007. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In Proceedings of Human Language Technology: the Conference of the North American Chapter of the Association for Computational Linguistics, pages 484-491

[2] Hua Wu and Haifeng Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. In Proceedings of 45th Annual Meeting of the Association for Computational Linguistics, pages 856-863.

[3] Trevor Cohn and Mirella Lapata. 2007. Machine Translation by Triangulation: Make Effective Use of Multi-Parallel Corpora. In Proceedings of 45th Annual Meeting of the Association for Computational Linguistics, pages 828-735.

[4] Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT:NAACL), pages 48-54, Edmonton, Canada, June.

[5] Tong Xiao, Jingbo Zhu, Hao Zhang, and Qiang Li. 2012. NiuTrans: An Open Source Toolkit for Phrase-based and Syntax-based Machine Translation. In Proceedings of ACL: System Demonstrations, pages 19-24, Jeju Island, Korea, July.

[6] Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In Proceedings of the 18th International Conference on Computational Linguistics, pages 1086-1090

[7] Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. Computer Speech & Language, 13:359-393.

[8] Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In Proceedings of ACL, pages 160-167, Sapporo, Japan, July.

[9] Kishore Papineni, Salim Roukos, Todd Ward and WeiJing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computation Linguistics, pages 311-319

[10] Jesús González-Rubio, Alfons Juan and Francisc Casacuberta. 2011. Minimum Bayes-risk System. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1268-1277

[11] Kevin Duh, Katsuhito Sudoh, Xianchao Wu, Hajime Tsukada and Masaaki Nagata. 2011. Generalized Minimum Bayes Risk System Combination. In Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1356-1360

[12] Kholy A E, Habash N, Leusch G, et al. Language Independent Connectivity Strength Features for Phrase Pivot Statistical Machine Translation[J]. Proc of Acl, 2013.

[13] Samira Tofighi Zahabi, Somayeh, Bakhshaei, Shahram Khadivi. 2013 . Using Context Vectors in Improving a Machine Translation System with Bridge Language. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 318-322.

[14] Xiaoning Zhu, Zhongjun He, Hua Wu, Haifeng Wang, et al. 2013. Improving Pivot-Based Statistical Machine Translation Using Random Walk. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 524-534.

[15] Xiaoning Zhu1, Zhongjun He, Hua Wu, et al. 2014. Improving Pivot-Based Statistical Machine Translation by Pivoting the Co-occurrence Count of Phrase Pairs. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1665-1675.

[16] Akiva Miura, Graham Neubig, Sakriani Sakti, et al. 2015. Improving Pivot Translation by Remembering the Pivot. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 573-577.

[17] Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita† et al. 2009. On the Importance of Pivot Language Selection for Statistical Machine Translation. Proceedings of NAACL HLT 2009: Short Papers, pages 221-224.