机器翻译服务变形测试的蒙特卡罗方法

丹尼尔·佩苏
卧龙岗计算与信息技术大学
卧龙岗，新南威尔士州2522，澳大利亚

智泉周*
计算机和信息技术网络安全和密码学院
卧龙岗大学，新南威尔士
2522，澳大利亚

京凤珍
卧龙岗计算与信息技术大学
卧龙岗，新南威尔士州2522，澳大利亚

戴夫·托威
宁波诺丁汉大学计算机科学学院
宁波315100，中国

摘要

随着机器翻译服务的日益普及，能够评估它们的质量变得越来越重要。 然而，测试oracle问题使得难以进行自动化测试。 在本文中，我们提出了一种蒙特卡罗方法，结合变质检验，以克服甲骨文问题。 使用这种方法，我们评估了三种流行的机器翻译服务的质量-即谷歌翻译、微软翻译和有道翻译。 我们将源语言设置为英语，目标语言包括汉语、法语、日语、韩语、葡萄牙语、俄语、西班牙语和瑞典语。 采用356个阶乘设计，收集和分析了33600个观测样本(共涉及100800个实际翻译。 基于这些数据，我们的模型发现Google翻译是所考虑的每一种目标语言的最佳(就所使用的变构关系而言。 印度的一个趋势-还确定了产生更好的结果的欧洲语言。

CCS概念

• 软件及其工程→经验软件验证；

关键词

机器翻译质量，甲骨文问题，变构测试，蒙特卡罗方法，自然语言

ACM 参考文献

格式：丹尼尔·佩苏，智泉周，景丰镇，戴夫·托维。 2018. 机器翻译服务变形测试的蒙特卡罗方法。 在MET '18：MET'18：IEEE/ACM国际变形测试讲习班，*2018年5月27日，瑞典哥德堡。* ACM，纽约，纽约，美国，8页。 https://doi.org/10.1145/3193977.3193980

* 所有信件都应寄给智泉周。 电子邮件：志泉@uow.edu.au

允许将本作品的所有或部分的数字或硬拷贝用于个人或课堂使用，但不收取费用，前提是副本不是为了利润或商业优势而制作或分发的，副本应承担本通知和第一页的全部引文。 必须尊重ACM以外的其他人拥有的这项工作的组件的版权。 允许用信用抽象。 若要以其他方式复制或重新发布、在服务器上发布或重新分发到列表，需要事先特定的权限和/或费用。 请求权限permissions@acm.org。

梅特18，2018年5月27日，瑞典哥德堡

©2018年计算机械协会。 ACM IS BN978-1-4503-5729-6/18/05.$15.00

https://doi.org/10.1145/3193977.3193980

1 导言

机器翻译是一种流行的应用程序，它解决了将文本从源语言自动翻译成目标语言的非常直接的需要。 如今，互联网上有很多免费翻译软件，比如谷歌翻译和微软的必应翻译器。 大型信息技术(IT)公司正在聘请研究人员和工程师生产自己的机器翻译产品，机器翻译市场[13]迅速增长。 机器翻译的新应用不断涌现，如移动设备上的实时自动语音翻译、跨语言信息检索和跨语言情感分析、自动字幕和字幕等[13]。

机器翻译质量的评价通常涉及人的干预和判断。 这是因为自动评估自然很难，因为没有测试甲骨文[2]。 一般来说，人力评估员的人工评估既昂贵又主观[19]。 因此，在本文中，我们考虑如何在没有人力评估员的情况下实现自动评估。

变形测试[5,7]被广泛认为是一种测试范式，可以有效地解决oracle问题[2,6,15]。 因此，在本研究中，我们探讨了在没有有形测试甲骨文的情况下，将变构测试应用于机器翻译服务的可能性和有效性。 更具体地说，我们提出以下研究问题：

- RQ1：在没有测试甲骨文（如人类评估员或目标语言中的同等文本)的情况下，是否可以将变构测试应用于测试机器翻译服务）？ RQ2：如果RQ1的答案是肯定的，
- 那么我们的方法在多大程度上可以区分变质测试框架中的好翻译服务和差翻译服务？

本文的其余部分组织如下：第二节描述了我们的变质关系和测试方法。 本节涉及RQ1。

为了解决RQ2问题，我们应用我们的方法来测试三种流行的机器翻译服务(谷歌翻译、微软翻译器，

和有道翻译)，并排名他们的一般表现在一些目标语言，以英语为来源（起源）语言。 目标语言是汉语、日语、韩语、法语、俄语、葡萄牙语、西班牙语和瑞典语。 第三节介绍了实验的设计，第四节分析了实验结果。 第5节包括进一步的讨论，第6节结束了论文。

2 我们的变质关系和测试方法

变质试验中最关键的任务是确定合适的变质关系(MRS)[6]。 MR是预期软件功能的必要属性。 它是被测试软件(SUT)多次执行的输入和输出之间的关系)。

对于机器翻译软件来说，最明显的MR可能是所谓的往返翻译(RT T)[1,10,16,17]：取一个初始字符串S，并将字符串从目标语言返回到原始语言进行双向翻译，从而导致 s'。 然后比较两个字符串S和S'，来评估它们的相似性。 从直觉上看，更高的相似性应该表明更好的翻译质量。 虽然RTT具有直观的吸引力，但由于其内在的局限性而受到批评：它不是测试一个系统，而是两个系统：前向翻译(FT)和后向翻译(BT)。 尽管有这一限制，一些研究人员报告说RTT可能是有用的。 例如，Aiken和Park[1]说，“RTT并不完美，但也没有其他评估技术。 对于单个给定的句子，我们不能确定好的（或坏的）RTT是否表明FT是好的（或坏的），反之亦然。 但是，在较长的文本或多个语言对的长度上，RTT质量可能反映了所使用的系统的一般质量。“他们进一步声称：

> *此外，RTT是唯一可以使用的技术，当没有人流利的目标语言或同等文本是容易获得的。*

在本文中，我们提出了一种非RTT技术，它可以在不需要等效目标语言文本或熟练（流利）目标语言用户的情况下使用。

在我们的方法中，我们实现了一种评估每个翻译服务质量的单向方法。 我们的方法在目标语言域执行比较过程，而不引用源语言。 通过这样做，我们避免了RTT方法的潜在缺点，同时保持了一种完全自动的方法来执行这些评估。

2.1 我们的变形关系

我们MR的总体思想是，一个理想的、完全一致的译者在直接（从源语言到目标语言)或间接(从源语言到中间语言，然后从中间语言到目标语言)翻译时，应该给出相同的翻译结果）。

为了实现以英语为源语言的MR，在每次变构测试中，我们首先将一个英语句子P译成
[illegible]
接下来，使用蒙特卡罗方法测量翻译的质量（一致性）如下：我们随机选择一个中间体语言M，使M既不是源语言，也不是目标语言。 然后将英文句子P译成中间语言，给出P_m。 最后，P_m被翻译成目标语言L，给出P_I'.
由此，利用比较函数对两者进行比较翻译P_I和P_I' 在语言L的领域，而不是源语言，英语。 在一个完美的翻译下期望这两个结果是相等的，从而给出了变质关系：

$$p_I = p_I' \quad (1)$$

方程（1）将称为MR1。 这个过程的一个例子如图1所示，英语是源语言，汉语是目标语言，日语是（随机选择的）中间语言。 单个变构测试将调用翻译服务三次：（1）翻译句子$P_{英语}$从英语到汉语，产生$P_{中文}$翻译$P_{英语}$从英语到日语，产生$P_{日语}$；和（3）翻译$P_{日语}$ 从日语到汉语，产生$P_{c'\ 汉尼斯}$. 最后，$P_{中文}$和$P_{c'\ 汉尼斯}$ 比较相似性。 更高的相似性将表明更一致（因此更高的质量）翻译结果。 图1显示了九种语言，在我们的实验中，英语总是被用作源（起源）语言，其他八种语言在我们的蒙特卡罗方法中依次被用作目标语言和中间语言。

2.2 比较度量

为了对两种翻译P进行有意义的比较$_I$和P_I' 需要一组度量来度量这些结果的相似性。总共考虑了三个指标：

- Levenshtein Distance，使一个翻译与另一个[18]相同所需的字符操作（插入、删除、替换）的最小数量。 这表示为最长平移的字符长度的比值。
- 布莱，它使用n-gram来确定[12]两个句子的相似性。 [3]使用了来自NLTK库的实现。
- 余弦相似度，它将这两个句子矢量化，并计算它们之间的[9]。

每个度量在0到1之间产生一个结果，其中1表示一个完美的匹配，0表示两个完全不同的句子。 这些度量用于比较两个翻译的相似性，并记录这三个度量的平均值用于每次比较。这个平均值作为MR1的分数。

3 实验设计

所提出的方法是在Python编写的工具中实现的，版本3.5.2。[1]

3.1 样本生成

为了生成所需的大量样本，通过爬取Wikipedia随机选择了一个英语句子列表[2] 并在被存储到文件之前进行预扫描。 这个过程被重复，以获得1000个有效的测试句子。

[1] https://www.python.org
[2] https://en.wikipedia.org

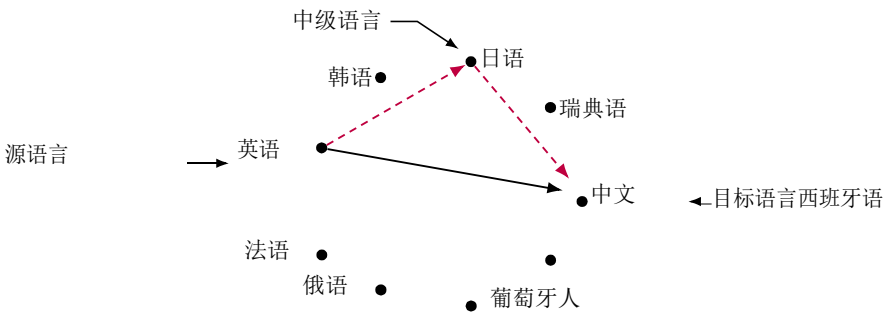

图1：翻译过程的说明。 黑线显示直接平移导致P$_I$彩色线显示导致P的路径$_I^*$. ×

3.2 实验设计

为了客观地分析结果，建立了一个统计模型，并按照其设计收集了数据。 在从方程（1)中获得比较分数时，我们考虑了两个变量或因素）：

(1) 用于生成P的翻译服务$_I$和P$_I^*$. *这个变量被称为翻译器，有三个选择，*
或层次：即谷歌翻译(https://translate.google. 微软翻译(https://translator.microsoft. 这是宾译者和有道翻译(http://fanyi.youdao.com)的权力)。 他们的API被调用来执行翻译。

(2) 用于翻译的中间语言M和目标语言L，它们被组合成一个变量：路径。 有八种目标语言（汉语、日语、韩语、法语、俄语、葡萄牙语、西班牙语和瑞典语）。对于每个目标语言，其他七种语言可以用作中间语言。 有，
因此，共有87对=56对独特的对组成
路径变量的级别。

这导致了一个356阶乘设计[8]与伴随的线性模型

$$s_{ijk}=\mu+T_i+p_j+(tp)_{ij}+\varepsilon_{ijk} . \quad (2)$$

在哪里，

- s_{ijk}是从JTH路径复制第i个翻译器的KTH复制的分数，
- μ是所有观测的全球平均值，
- t_i是第一译者和μ的平均分数之间的差异，称为第一译者的主要效果，P_jJTH路径的平均分数与μ之间的差
- 异，称为JTH路径的主要效应，
- $(tp)_{ij}$是组合效果的相互作用项
翻译与JTH路径，和
- ε_{ijk}是与每次观测相关的随机误差(或残差。 这解释了模型的预测和观测值之间的差异。

假设残差的值采取正态随机变量的形式，在数据拟合到模型后，均值为0，方差为常数。

有3个56=168个翻译路径组合。 这些治疗组合中的每一个都被复制了200次，每个复制都从1000个列表中随机选择一个新句子，并替换。 为了本实验的目的，每个句子都被系统地分配了一种中间语言，而不是随机选择中间语言。 通过这样做，我们可以确保每个治疗组合收到相同数量的复制，因此，我们有一个完整的阶乘设计[8]。

本设计[8]的观测单元和实验单元是对平移P$_I$和P* 。 共产生200个168个=33个600个观测值。 ×因为每个观察都涉及三个实际翻译(生成P$_I$ ，p$_I^*$ 以及中间翻译-
结果P$_m$)，本研究共产生33600=100800
实际翻译。 ×

我们已经在Zenodo网上提供了我们的数据集和测试结果：
http://doi.org/10.5281/zenodo.1194560.

4 实验结果

所有分析均使用统计软件R，版本进行
3.4.3[14]使用lsmeans包[11]中的方法。

4.1 模型诊断

根据方程（2）的模型对33,600个观测值进行了拟合，并通过验证每个假设都得到了满足来评估该模型的适用性。 正态性的统计检验，如Shapiro-Wilk检验和常数方差的检验，如Levene检验，由于这些检验具有如此大的样本大小而具有的膨胀功率，因此没有使用。

- 残差的独立性，每个观察都是随机选择的，因此是独立的。
- 残差的正态性，图2中的Q-Q图显示了一个合理的拟合，在负尾部略有偏离正态性。
- 残差的恒定方差，图2中的残差图显示了残差的系统变化的一些证据，与中间分数相比，较高和较低的拟合分数的变化较小。

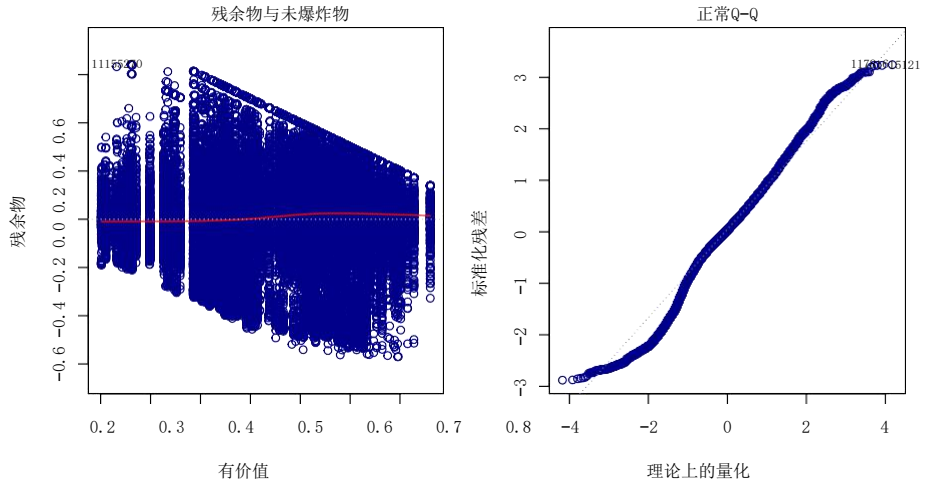

图2：残差图（左），以评估常数方差的假设。 这是由每个拟合值的点的恒定垂直扩展来表示的。 还产生了一条趋势线（红色）来识别任何系统的行为。 用Q-Q图（右）来评估正态性，这是由数据与对角线的拟合程度来表示的。

该模型合理地满足了每个假设，因此适合于分析数据。

4.2 模型的意义

为了检验方程（2）中主要效应之间的任何差异，进行了方差分析(ANOVA)，结果见表1。

在5%的显著性水平下，方差分析的结果表明，译者与路径之间的相互作用的影响，表示译者 径并以(TP)表示)$_{ij}$在方程（2）中，不等于所有级别(F（110，33432）=7。79，p<。0001)．这表明，主要影响不是简单的加性，因此不能[8]单独评估。 相反，我们将通过简单的效果来检查每个因素组合的估计。

为了简单起见，对具有公共目标语言的路径采取了平均简单效果，并在表2中报告了每个翻译-目标语言对的估计。 为了确定哪种翻译器对每种目标语言的表现最好，反之亦然，[4]进行了事后多次比较。

4.3 多重比较分析

在固定的目标语言中，每个译者的估计被称为译者的简单效果。 比较每一个简单的效果，译者可以从最高到最低。 进行多重比较分析(成对t检验)[4]以确定这些简单效果中哪些是显著不同的。 这些排名是由表2中[4]的齐次子集给出的。

同样，通过固定特定的翻译服务，每个目标语言的估计都是路径的简单效果（按目标语言分组）。 图3绘制了每个翻译器中这些效果的置信区间。 再次，通过表演事后多重比较分析，每种语言的排名可以识别。 这些排名是由表3中的齐次子集给出的。

总之，对跨路径组翻译的简单效果的分析表明，谷歌翻译每种目标语言的平均得分都明显高于其他两种服务。 微软笔译员在中国、日本、法国和俄罗斯的目标路径上的平均得分明显高于有道。 微软和有道在韩国、葡萄牙和西班牙的目标路径上的平均得分没有显著差异。 对于瑞典人来说，有道的平均分数比微软高得多。

对跨译者路径组的简单效果的分析发现，西班牙语和葡萄牙语是所有三个翻译中唯一出现在最高排名的两种目标语言。 另一方面，韩语、日语和汉语目标语言通常产生最低的平均分数，汉语目标语言总是出现在底部。

5 讨论

从这些结果中，我们发现，对于所有八种非英语目标语言，谷歌翻译是最一致的服务，最好满足变质关系MR1。 本研究的一个局限性是，虽然这种一致性的概念是直接测试的，但它不一定是翻译正确性的充分条件；然而，从用户的角度来看，翻译一致性在任何情况下都是一个可取的属性。

韩国语、日语和汉语目标语言被发现是表现最差的目标语言。 进一步的检查发现，BLEU和Cosine相似评分对这些语言的惩罚更大。

为了提高本研究的有效性，我们进行了一项小规模的随访研究。 我们拿了一套140直接的英汉

表1：译者和路径的双向方差分析。

变异的来源	平方和	自由程度	平均平方		F比值	p值
译者（主效）	94.46	2	47.23	1,199.1611	<	.0001
路径（主要效果）	790.71	55	14.38	365.0368	<	.0001
译者×路径（相互作用）	33.76	110	0.31	7.7927	<	.0001
剩余（错误）	1,316.69	33,432	0.04			
共计	2,235.62	33,599				

每一行的P值对应于零假设，即变异源的每个级别具有相等的效果。

表2：译者和齐次子集的简单效应。

目标语言	笔译员	估计数	小组	
中文	谷歌	0.3414364a		
	微软	0.2609551	b	
	有道	0.2200295		c
日语	谷歌	0.4308912a		
	微软	0.4035918	b	
	有道	0.3848805		c
韩语	谷歌有道	0.6236494a		
	微软	0.5185094	b	
		0.5168478	b	
法语	谷歌	0.7070268a		
	微软	0.5810083	b	
	有道	0.5611871		c
俄语	谷歌	0.6819666a		
	微软	0.5519567	b	
	有道	0.5338486		c
葡萄牙语	谷歌	0.7209871a		
	有道	0.5894090	b	
	微软	0.5723823	b	
西班牙语	谷歌	0.7007260a		
	微软	0.5935882	b	
	有道	0.5902570	b	
瑞典语	谷歌	0.7094092a		
	有道	0.5885621	b	
	微软	0.5688555		c

每个估算的标准误差为0.005303906。 没有出现在同一组中的估计值有显着性差异。 这8组3个比较中的每一个都是在Tukey调整的家庭误差率为5%的情况下进行的。

翻译原始实验结果，如第4节所示。 这一套包括70个谷歌翻译和70个微软翻译。 然后，我们邀请了一位母语为汉语的用户，他们生活和互动在英语语言媒体（在澳大利亚），手动评估翻译质量。 每个翻译都使用以下标准进行评分：

- 如果翻译完全错误，
- 如果翻译有意义但很差，
- 如果翻译有意义，但有小错误，和
- 如果这是一个完美的翻译。

人类评估分数汇总在表4中，显示谷歌（微软)收到了32(8）3s，20六（二十三）2s和十二（三十九）1s。 这些分数表明，谷歌在英汉翻译方面优于微软。 这个人类评估结果与第4节中提出的机器评估结果是一致的，因为谷歌表现最好。

6 结论和今后的工作

机器翻译的质量很难通过自动化手段来评估。 在本文中，我们提出了两个研究问题：在没有甲骨文的情况下，甲骨文的情况下应用，例如人类评估员或目标语言中的等效文本？ 此外，变形测试在多大程度上可以区分好的和差的翻译服务？ 本研究的结果

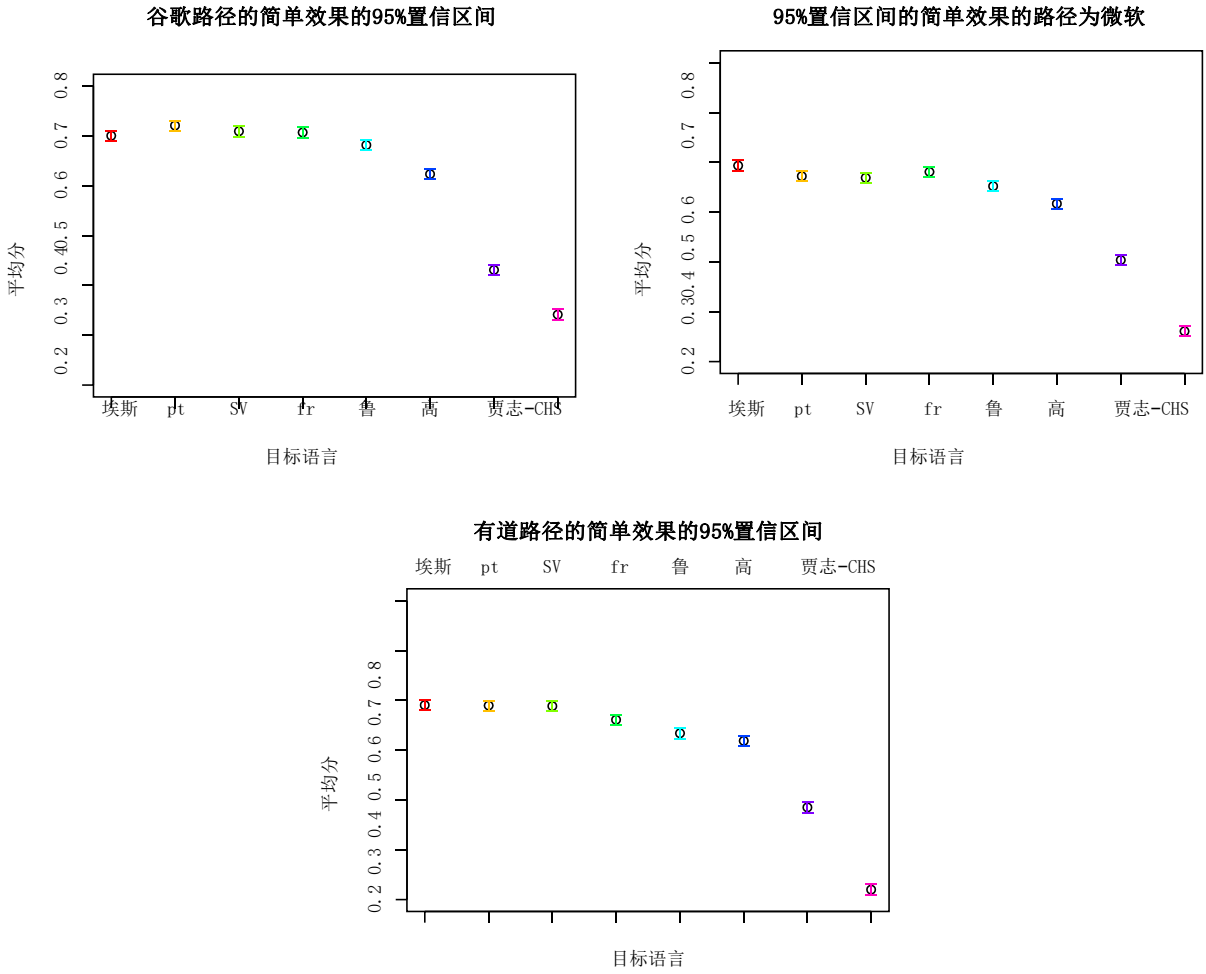

图3：按通用目标语言分组的路径在每个翻译器上的简单效果。 es：西班牙语，pt：葡萄牙语，SV：瑞典语，fr：法语，ru：俄语，ko：韩语，ja：日语，zh-CHS：简体中文。

对这两个研究问题提供肯定的答复。 我们提出了一种新的方法来自动评估机器翻译服务使用简单的MR，以避免往返翻译(RTT) [3]。 该方法的实证结果包括33,600个观察结果（涉及谷歌、微软和有道翻译服务为一种源语言(英语)和八种目标语言产生的总共100,800个实际机器翻译输出）。 通过对这些结果进行统计分析，我们能够客观地确定哪些服务最能满足MR（谷歌翻译），哪些领域面临挑战(翻译成亚洲语言）。

[3] 尽管如此，一项进一步的研究表明，我们的结果和RTT的结果往往是一致的。 然而，对RTT结果的进一步讨论超出了本文的范围。

为了提高我们的发现的有效性，我们进行了一项小规模的后续研究，其中包括使用一名人类评估员、一种目标语言（中文)和两种翻译服务(谷歌和微软）。 发现人的评价结果与前面章节报道的机器评价结果一致。 平均而言，人类评估员要花30多秒来评估一篇翻译。 为了执行相同的任务，我们的自动化测试工具需要不到两秒。 这种比较表明，我们的方法具有很高的成本效益。 我们的发现的外部有效性可以通过使用不同的源语言和维基百科以外的不同样本源进行更大规模的实验来增强。 在今后的研究中，我们将进一步研究机器评价分数与人类评价分数之间的相关性。 我们也计划

表3：路径和齐次子集的简单效应。

	目标语言	估计数	葡萄			
谷歌	牙语集团	0.7209871a				
	瑞典语	0.7094092	a			
	法语	0.7070268	a			
	西班牙语	0.7007260	a	b		
	俄语	0.6819666	b			
	韩语	0.6236494		c		
	日语	0.4308912			d	
	中文	0.3414364				e
微软	西班牙语	0.5935882a				
	法语	0.5810083a b				
	葡萄牙语	0.5723823A B C瑞典				
	语	0.5688555	b			
	c					
	俄语	0.5519567		c		
	韩语	0.5168478			d	
	日语	0.4035918				e
	中文	0.2609551				f
有道	西班牙语	0.5902570	a			
	葡萄牙语	0.5894090a				
	瑞典语	0.5885621	a			
	法语	0.5611871	b			
	俄语	0.5338486		c		
	韩语	0.5185094		c		
	日语	0.3848805			d	
	中文	0.2200295				e

每个估算的标准误差为0.005303906。 没有出现在同一组中的估计值有显着性差异。 这3组28个比较中的每一组都是在Tukey调整的家庭误差率为5%的情况下进行的。

表4：140份汉译英直接评估分数摘要

服务	0	1 3	2	共计	
谷歌	0	12	26	3270	
微软	0	39	23	8	70
共计	0	51	49	40140	

调查我们的发现对各种应用领域的影响，如跨语言信息检索和跨语言情感分析。

感谢

这项工作得到了澳大利亚研究理事会的联系赠款(项目ID：LP160101691)的部分支持)。 我们也要感谢苏州英视云信息技术有限公司支持这项研究。 我们感谢卧龙岗大学的肯尼斯·罗素对这项工作的宝贵意见。 我们要感谢卧龙岗大学的下列学生在这项工作的初步研究中为部分实施作出了贡献：基兰·麦克雷、丹尼尔·巴恩斯、王世新和博阳燕。

参考资料

[1] Milam Aiken和Mina Park。 2010. 双程翻译对MT评价的有效性。 *翻译期刊14，1（2010）。* http://translationjournal.net/journal/51reverse.htm

[2] Earl T。 巴尔、马克·哈曼、菲尔·麦克明恩、穆扎米尔·沙巴兹和申宇。 2015. 软件测试中的甲骨文问题：一项调查。 *IEEE交易*

软件工程41，5（2015），507-525。
[3] 史蒂文·伯德，伊万·克莱恩和爱德华·洛珀。 2009. *自然语言处理与Python-分析文本与自然语言工具包。* 奥雷利媒体。 http://www.nltk.org/
[4] S.卡默和史旺森。 1973. 蒙特卡罗方法对十个两两多重比较程序的评价。 *j. Amer。 是个政治家。 协会。* 68，341（1973），66－74. https://doi.org/10.2307/2284140
[5] t. y. 陈，S。 c. 张和S。 m. 尤。 1998. 变形测试：生成下一个测试用例的新方法。 技术报告HKUST-CS98-01。 香港科技大学计算机科学系。
[6] 陈宗岳，郭飞成，刘怀，朴龙鹏，戴夫·托威，T。 h. 策，和志全周。 2018. 变形测试：挑战和机遇的回顾。 ACM计算调查51，1（2018），4：1-4：27。
[7] t. y. 陈，T。 h. 谢和周泽强。 2003. 基于故障的测试，不需要支架。 *信息和软件技术45，1（2003），1-9。*
[8] 大卫·罗克斯比·考克斯和南希·里德。 2000. *实验设计理论。 查普曼和霍尔/CRC。*
[9] 安娜·黄。 2008. 文本文档聚类的相似性度量。 在第六届新西兰计算机科学研究学生会议记录(NZCSRSC，2008年)。 克赖斯特彻奇，新西兰，49-56。
[10] 菲利普·科恩和克里斯多夫·蒙兹。 2006. 欧洲语言之间机器翻译的手动和自动评估。 *统计机器翻译讲习班记录(StatMT‘06)。 计算语言学协会，102-121。*
[11] Russell V。lenth。 2016.最小二乘均值：R包lsmeans。 *统计软件杂志69，1（2016），1-33。*

[12] 基肖尔·帕皮涅尼、萨利姆·鲁科斯、托德·沃德和朱伟静。 2002. 机器翻译的自动评价方法。 在第40届计算语言学协会年会论文集(ACL'02)。计算语言学协会，斯特罗斯堡，宾夕法尼亚州，美国，311-318。 https://doi.org/10.3115/1073083.1073135

[13] 蒂埃里·波贝奥。 2017. *机器翻译。 麻省理工学院出版社。*

[14] R核心小组。 2017. *R：统计计算的语言和环境。* R统计计算基金会，奥地利维也纳。 https://www.R-project. org/

[15] Sergio Segura、Gordon Fraser、Ana B.Sanchez和Antonio Ruiz-Cortes。 2016. 变质试验调查。 *IEEE软件工程交易42，9（2016），805-824。*

[16] Tomohiro Shigenobu。 2007. 跨文化交际背译的评价与可用性。 在第二次可用性和国际化国际会议记录，计算机科学讲座笔记，第4560卷。 斯普林格-维拉格，259-265。

[17] H.萨默斯。 2005. 往返翻译：它有什么好处？。 澳大利亚语言技术研讨会论文集。 127 - 133.

[18] 张胜南，严虎，卞光荣。 2017. 基于Levenshtein距离的字符串相似度算法研究。 在IEEE第二届高级信息技术，电子和自动化控制会议(IAEAC)。 2247 - 2251. https://doi.org/10.1109/IAEAC.2017.8054419

[19] 志全周，少文祥，和宋月辰。 2016. 软件质量评估的变形测试：搜索引擎的研究。 *IEEE软件工程交易42，3（2016），264-284。*