A Monte Carlo Method for Metamorphic Testing of Machine Translation Services

Preprint · May 2018

DOI: 10.1145/3193977.3193980

4 authors, including:

Zhi Quan Zhou
University of Wollongong
51 PUBLICATIONS **831** CITATIONS

SEE PROFILE

Dave Towey
University of Nottingham, Ningbo Campus
94 PUBLICATIONS **494** CITATIONS

SEE PROFILE

A Monte Carlo Method for Metamorphic Testing of Machine Translation Services

Daniel Pesu
School of Computing and Information Technology
University of Wollongong
Wollongong, NSW 2522, Australia

Zhi Quan Zhou*
Institute of Cybersecurity and Cryptology
School of Computing and Information Technology
University of Wollongong
Wollongong, NSW 2522, Australia

Jingfeng Zhen
School of Computing and Information Technology
University of Wollongong
Wollongong, NSW 2522, Australia

Dave Towey
School of Computer Science
University of Nottingham Ningbo China
Ningbo 315100, China

ABSTRACT

With the growing popularity of machine translation services, it has become increasingly important to be able to assess their quality. However, the test oracle problem makes it difficult to conduct automated testing. In this paper, we propose a Monte Carlo method, in combination with metamorphic testing, to overcome the oracle problem. Using this method, we assessed the quality of three popular machine translation services — namely, Google Translate, Microsoft Translator, and Youdao Translate. We set the source language to be English, and the target languages included Chinese, French, Japanese, Korean, Portuguese, Russian, Spanish, and Swedish. A sample of 33,600 observations (involving a total of 100,800 actual translations) was collected and analyzed using a 3 × 56 factorial design. Based on this data, our model found Google Translate to be the best (in terms of the metamorphic relation used) for each and every target language considered. A trend for Indo-European languages producing better results was also identified.

CCS CONCEPTS

• **Software and its engineering → Empirical software validation**;

KEYWORDS

Machine translation quality, oracle problem, metamorphic testing, Monte Carlo method, natural languages

ACM Reference Format:
Daniel Pesu, Zhi Quan Zhou, Jingfeng Zhen, and Dave Towey. 2018. A Monte Carlo Method for Metamorphic Testing of Machine Translation Services. In *ICSE MET '18*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3193977.3193980

*All correspondence should be addressed to Zhi Quan Zhou. Email: zhiquan@uow.edu.au

This is a preprint. In Proceedings of the IEEE/ACM 3rd International Workshop on Metamorphic Testing (MET '18), in conjunction with the 40th International Conference on Software Engineering (ICSE '18). May 27, 2018, Gothenburg, Sweden.

https://doi.org/10.1145/3193977.3193980

1 INTRODUCTION

Machine translation is a popular application that addresses a very direct need for automatically translating texts from a source language into a target language. Nowadays, a lot of free translation software is available on the Internet, such as Google Translate and Microsoft's Bing Translator. Large information technology (IT) companies are hiring researchers and engineers to produce their own machine translation products, and the machine translation market is growing rapidly [13]. New applications of machine translation continue to emerge, such as live automated speech translation on mobile devices, cross-language information retrieval and cross-language sentiment analysis, automatic subtitling and captioning, and so on [13].

Evaluation of machine translation quality normally involves human intervention and judgment. This is because automatic assessment is naturally difficult, due to the lack of a *test oracle* [2]. Generally speaking, manual assessment by a human assessor is both expensive and subjective [19]. In this paper, therefore, we consider how to achieve automatic assessment without a human assessor.

Metamorphic testing [5, 7] is widely recognized as a testing paradigm that can effectively address the oracle problem [2, 6, 15]. In this study, therefore, we explore the possibility and effectiveness of applying metamorphic testing to test machine translation services in the absence of a tangible test oracle. More specifically, we raise the following research questions:

- RQ1: Can metamorphic testing be applied to test machine translation services in the absence of a test oracle (such as a human assessor or an equivalent text in the target language)?
- RQ2: If the answer to RQ1 is affirmative, then to what degree can our approach distinguish between good and poor translation services in the context of the metamorphic testing framework?

The rest of this paper is organized as follows: Section 2 describes our *metamorphic relation* and our testing method. This section addresses RQ1.

To address RQ2, we apply our method to test three popular machine translation services (Google Translate, Microsoft Translator,

and Youdao Translate), and rank their general performance on a number of target languages with English as the source (origin) language. The target languages are Chinese, Japanese, Korean, French, Russian, Portuguese, Spanish, and Swedish. Section 3 presents the design of the experiments, and Section 4 analyzes the experimental results. Section 5 includes further discussion and Section 6 concludes the paper.

2 OUR METAMORPHIC RELATION AND TESTING METHOD

The most critical task in metamorphic testing is the identification of suitable *metamorphic relations* (MRs) [6]. An MR is a necessary property of the intended software's functionality. It is a relation among the inputs and outputs of multiple executions of the software under test (SUT).

For machine translation software, the most obvious MR is probably the so-called *round-trip translation* (RTT) [1, 10, 16, 17]: Take an initial string S and perform a two-way translation of the string from a target language back to the original language, resulting in S'. Then a comparison is made between the two strings, S and S', to assess their similarity. Intuitively, higher similarity should indicate better translation quality. Though intuitively attractive, RTT has been criticized for its intrinsic limitation: It is not testing one system, but two systems: the forward translation (FT) and the back translation (BT). Despite this limitation, some researchers reported that RTT could be useful. For example, Aiken and Park [1] stated that "RTT is not perfect, but no other evaluation technique is, either. For a single given sentence, we cannot know for sure if a good (or bad) RTT indicates that the FT was good (or bad) or vice versa. But, over the length of a longer text or multiple language pairs, RTT quality might reflect the general quality of the system used." They further claimed that:

> *In addition, RTT is **the only** technique that can be used when no human fluent in the target language or equivalent text is readily available.*

In this paper, we propose a **non-RTT** technique that can be used without the need for an equivalent target language text, or proficient (fluent) target language user.

In our approach, we implement a one-way method for evaluating the quality of each translation service. Our method performs the comparison process at the target language domain without referring back to the source language. By doing so, we avoid the potential shortcomings of the RTT methods, while maintaining an entirely automatic method for performing these evaluations.

2.1 Our Metamorphic Relation

The general idea of our MR is that an *ideal*, *perfectly consistent* translator should give the same translation results when translating either directly (from a *source* language to a *target* language) or indirectly (from the source language to an *intermediate* language and then from the intermediate language to the target language).

To implement this MR with English as the source language, in each metamorphic test we first translate an English sentence P into a target language L. Let P_L denote this direct translation result. Next, a Monte Carlo method is used to measure the quality (consistency) of translation as follows: We randomly select an intermediate language M such that M is neither the source language nor the target language. The English sentence P is then translated into the intermediate language, giving P_M. Finally, P_M is translated into the target language L, giving P'_L.

From this, a comparison function is used to compare the two translations P_L and P'_L in the domain of the language L rather than the source language, English. Under a *perfect* translator we would expect these two results to be equal, thus giving the metamorphic relation:

$$P_L = P'_L \tag{1}$$

Equation (1) will be referred to as **MR1** hereafter. An example of this process is illustrated in Figure 1, with English being the source language, Chinese being the target language, and Japanese being the (randomly chosen) intermediate language. A single metamorphic test will invoke the translation service three times: (1) translate a sentence P_{English} from English to Chinese, yielding P_{Chinese}; (2) translate P_{English} from English to Japanese, yielding P_{Japanese}; and (3) translate P_{Japanese} from Japanese to Chinese, yielding P'_{Chinese}. Finally, P_{Chinese} and P'_{Chinese} are compared for similarity. Higher similarity would indicate more consistent (hence higher quality) translation results. Figure 1 shows nine languages, where English was always used as the source (origin) language in our experiments, and the other eight languages were used as target and intermediate languages in turn in our Monte Carlo approach.

2.2 Comparison Metrics

In order to make meaningful comparisons between the two translations P_L and P'_L, a set of metrics are needed to measure the similarity of these results. In total, three metrics are considered:

- **Levenshtein Distance**, the minimum number of character-wise operations (insert, delete, replace) needed to make one translation identical to the other [18]. This is expressed as a ratio of the character length of the longest translation.
- **BLEU**, which uses n-grams to determine the similarity of two sentences [12]. The implementation from the NLTK library was used [3].
- **Cosine Similarity**, which vectorizes the two sentences and calculates the angle between them [9].

Each metric produces a result between 0 and 1, where 1 represents a perfect match, and 0 represents two entirely different sentences. These metrics are used to compare the similarity of the two translations, and the average of these three metrics is recorded for each comparison. This average is taken as the *score* for MR1.

3 DESIGN OF EXPERIMENTS

The proposed method was implemented in a tool written in Python, version 3.5.2. [1]

3.1 Sample Generation

To generate the large number of samples required, a list of English sentences was randomly selected by crawling Wikipedia [2] and pre-scanned before being stored into a file. This process was repeated to obtain 1,000 valid test sentences.

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

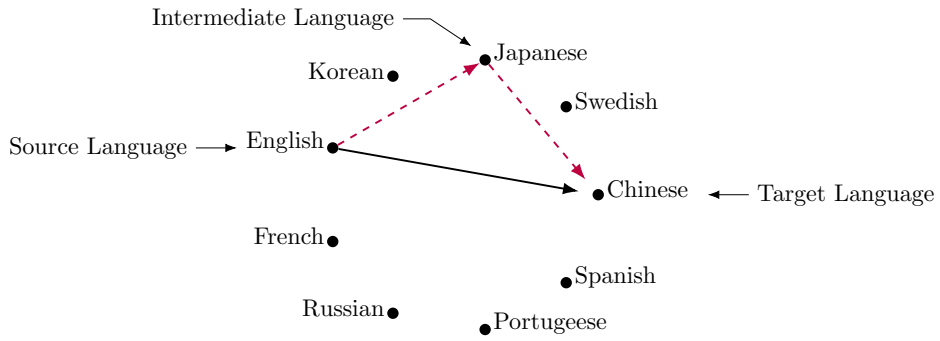

Figure 1: Illustration of the translation procedure. The black line shows the direct translation resulting in P_L and the coloured line shows the path resulting in P'_L.

3.2 Experimental Design

To objectively analyze the results, a statistical model was developed and the data was collected in accordance with its design. There are two variables, or *factors*, which we consider when obtaining the comparison score from Equation (1):

(1) The translation service used to generate P_L and P'_L. This variable is referred to as the *translator* and has three choices, or *levels*: namely, Google Translate (https://translate.google.com.au), Microsoft Translator (https://translator.microsoft.com), which powers the Bing Translator, and Youdao Translate (http://fanyi.youdao.com). Their APIs were called to perform the translations.
(2) The intermediate language, M, and target language, L, used for the translation, which are combined into a single variable: *path*. There are eight target languages (Chinese, Japanese, Korean, French, Russian, Portuguese, Spanish, and Swedish). For each target language, the seven other languages can be used to serve as the intermediate language. There are, therefore, a total of $8 \times 7 = 56$ unique pairs which make up the levels of the path variable.

This gives rise to a 3×56 *factorial design* [8] with the accompanying linear model

$$S_{ijk} = \mu + T_i + P_j + (TP)_{ij} + \varepsilon_{ijk}. \quad (2)$$

where,

- S_{ijk} is the score for the kth replication of the ith translator from the jth path,
- μ is the global mean for all observations,
- T_i is the difference between the mean score for the ith translator and μ, referred to as the *main effect* of the ith translator,
- P_j the difference between the mean score for the jth path and μ, referred to as the *main effect* of the jth path,
- $(TP)_{ij}$ is an *interaction* term for the effect of combining the ith translator with the jth path, and
- ε_{ijk} is the *random error* (or *residual*) associated with each observation. This accounts for the difference between the model's prediction and the observed value.

The values of the residuals are assumed to take the form of a normal random variable with mean 0 and constant variance after the data has been fitted to the model.

There are $3 \times 56 = 168$ translator-path combinations. Each of these treatment combinations was replicated 200 times, with a new sentence being randomly selected from the list of 1,000, with replacement, for each replication. For the purpose of this experiment, each sentence was systematically allocated an intermediate language, rather than randomly selecting the intermediate language. By doing so, we can ensure that each treatment combination receives the same number of replications and, hence, that we have a *complete* factorial design [8].

The *observational units* and *experimental units* [8] in this design are pairs of translations P_L and P'_L. A total of $200 \times 168 = 33,600$ observations were generated. As each observation involved three actual translations (to generate P_L, P'_L, and the intermediate translation result P_M), this study generated a total of $3 \times 33,600 = 100,800$ actual translations.

We have made our data set and test results available online at Zenodo: http://doi.org/10.5281/zenodo.1194560.

4 RESULTS OF EXPERIMENTS

All analysis was performed using the statistical software R, version 3.4.3 [14] using methods from the *lsmeans* package [11].

4.1 Model Diagnostics

The 33,600 observations were fitted according to the model of Equation (2), and the suitability of this model was assessed by verifying that each of the assumptions has been met. Statistical tests of normality, such as the Shapiro-Wilk test, and tests of constant variance, such as Levene's test, were not used due to the inflated power these tests will have from having such large sample sizes.

- *Independence of residuals*, each observation was randomly selected and therefore independent.
- *Normality of residuals*, the Q-Q plot in Figure 2 shows a reasonable fit with a slight departure from normality in the negative tail.
- *Constant variance of residuals*, the residual plot in Figure 2 shows some evidence of systematic variation in residuals with less variation in higher and lower fitted scores compared to middle scores.

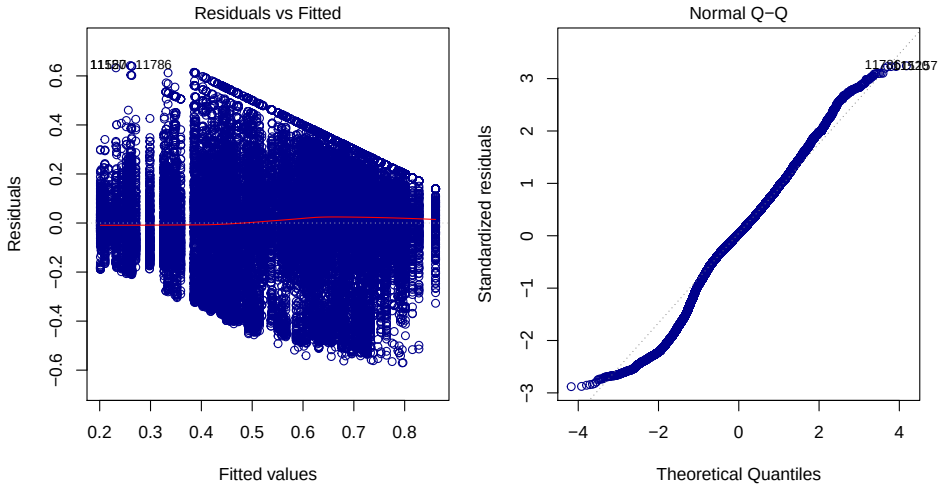

Figure 2: Residual plot (left) to assess the assumptions of constant variance. This is indicated by a constant vertical spread of points for each fitted value. A trend line (red) is also produced to identify any systematic behaviour. A Q-Q plot (right) is used to assess normality which is indicated by how well the data fit the diagonal line.

The model reasonably satisfies each of the assumptions and therefore is appropriate for analyzing the data.

4.2 Significance of the Model

To test for any differences between the main effects in Equation (2), a Two-Way Analysis of Variance (ANOVA) was conducted, with the results given in Table 1.

At the 5% level of significance, the results from the ANOVA suggest that the effects of the interaction between translator and path, denoted Translator × Path and represented by $(TP)_{ij}$ in Equation (2), is not equal for all levels ($F(110, 33432) = 7.79$, $p < .0001$). This indicates that the *main effects* are not simply additive and, therefore, cannot be assessed separately [8]. Instead, we will examine estimates of each combination of factors through *simple effects*.

For simplicity, the average simple effect for paths with common target languages is taken and estimates for each translator-target language pair are reported in Table 2. To determine which translator performs best for each target language and vice versa, post-hoc multiple comparisons were made [4].

4.3 Multiple Comparisons Analysis

The estimates for each of the translators within a fixed target language are known as the *simple effects* of translator. Comparing each of the simple effects, the translators can be ranked from highest to lowest. Multiple comparison analysis (pair-wise t-tests) [4] was carried out to determine which of these simple effects are significantly different from one another. These rankings are given by the homogeneous subsets [4] in Table 2.

Similarly, by fixing a particular translation service, the estimates for each of the target languages are the simple effects of path (grouped by target language). Confidence intervals for these effects within each translator are plotted in Figure 3. Again, by performing post-hoc multiple comparisons analysis, rankings for each language can be identified. These rankings are given by the homogeneous subsets in Table 3.

In summary, the analysis of the simple effects of translators across path groups revealed that Google Translate produced a significantly higher average score than the other two services for every target language. Microsoft Translator had a significantly higher average score than Youdao in paths with Chinese, Japanese, French, and Russian targets. There was no significant difference between the average scores for Microsoft and Youdao in paths with Korean, Portuguese, and Spanish targets. For Swedish, Youdao had a significantly higher average score than Microsoft.

The analysis of the simple effects of path groups across translators found that Spanish and Portuguese were the only two target languages appearing in the top rank for all three translators. At the other end, Korean, Japanese, and Chinese target languages commonly produced the lowest average scores, with the Chinese target language always appearing at the bottom.

5 DISCUSSION

From these results we have found that, for all eight of the non-English target languages considered, Google Translate is the most consistent service, best satisfying the metamorphic relation MR1. A limitation of this study is that, although this notion of consistency is what was directly being tested, it is not necessarily a sufficient condition for correctness of translation; however, translation consistency is in any case a desirable property from the user's perspective.

The Korean, Japanese, and Chinese target languages were found to be the worst performing target languages. Further inspection found that the BLEU and Cosine similarity scores were more penalizing to these languages.

To enhance the validity of this research, we conducted a small-scale follow-up study. We took a set of 140 *direct* English-Chinese

Table 1: Two-way ANOVA for translator and path.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F ratio	p-value
Translator (main effect)	94.46	2	47.23	1,199.1611	$< .0001$
Path (main effect)	790.71	55	14.38	365.0368	$< .0001$
Translator × Path (interaction)	33.76	110	0.31	7.7927	$< .0001$
Residual (Error)	1,316.69	33,432	0.04		
Total	2,235.62	33,599			

***P*-values for each row correspond to the null hypothesis that each level of the source of variation has an equal effect.**

Table 2: Simple effects of translator and homogeneous subsets.

Target Language	Translator	Estimate	Group		
Chinese	Google	0.3414364	A		
	Microsoft	0.2609551		B	
	Youdao	0.2200295			C
Japanese	Google	0.4308912	A		
	Microsoft	0.4035918		B	
	Youdao	0.3848805			C
Korean	Google	0.6236494	A		
	Youdao	0.5185094		B	
	Microsoft	0.5168478		B	
French	Google	0.7070268	A		
	Microsoft	0.5810083		B	
	Youdao	0.5611871			C
Russian	Google	0.6819666	A		
	Microsoft	0.5519567		B	
	Youdao	0.5338486			C
Portuguese	Google	0.7209871	A		
	Youdao	0.5894090		B	
	Microsoft	0.5723823		B	
Spanish	Google	0.7007260	A		
	Microsoft	0.5935882		B	
	Youdao	0.5902570		B	
Swedish	Google	0.7094092	A		
	Youdao	0.5885621		B	
	Microsoft	0.5688555			C

The standard error for each estimate is 0.005303906. Estimates not appearing in the same group are significantly different. Each of the 8 sets of 3 comparisons were made under a Tukey-adjusted family-wise error rate of 5% for each set.

translations from the original experimental results as presented in Section 4. This set consisted of 70 Google translations and 70 Microsoft translations. We then invited a native Chinese language user, who lives and interacts in an English language medium (in Australia), to manually assess the translation quality. Each and every translation was scored using the following criteria:

- 0 if the translation is completely wrong,
- 1 if the translation makes some sense but is poor,
- 2 if the translation makes sense but has minor errors, and
- 3 if this is a perfect translation.

The human assessment scores are summarized in Table 4, which shows that Google (Microsoft) received thirty two (eight) 3s, twenty six (twenty three) 2s, and twelve (thirty nine) 1s. These scores suggest that Google outperformed Microsoft in the English-Chinese translations. This human evaluation outcome is consistent with the machine evaluation outcome presented in Section 4 in that Google performed the best.

6 CONCLUSION AND FUTURE WORK

The quality of machine translation is difficult to assess by automated means. In this paper, we raised two research questions: Can metamorphic testing be applied in the absence of an oracle, such as a human assessor or an equivalent text in the target language? Further, to what degree can metamorphic testing distinguish between good and poor translation services? The results of this research

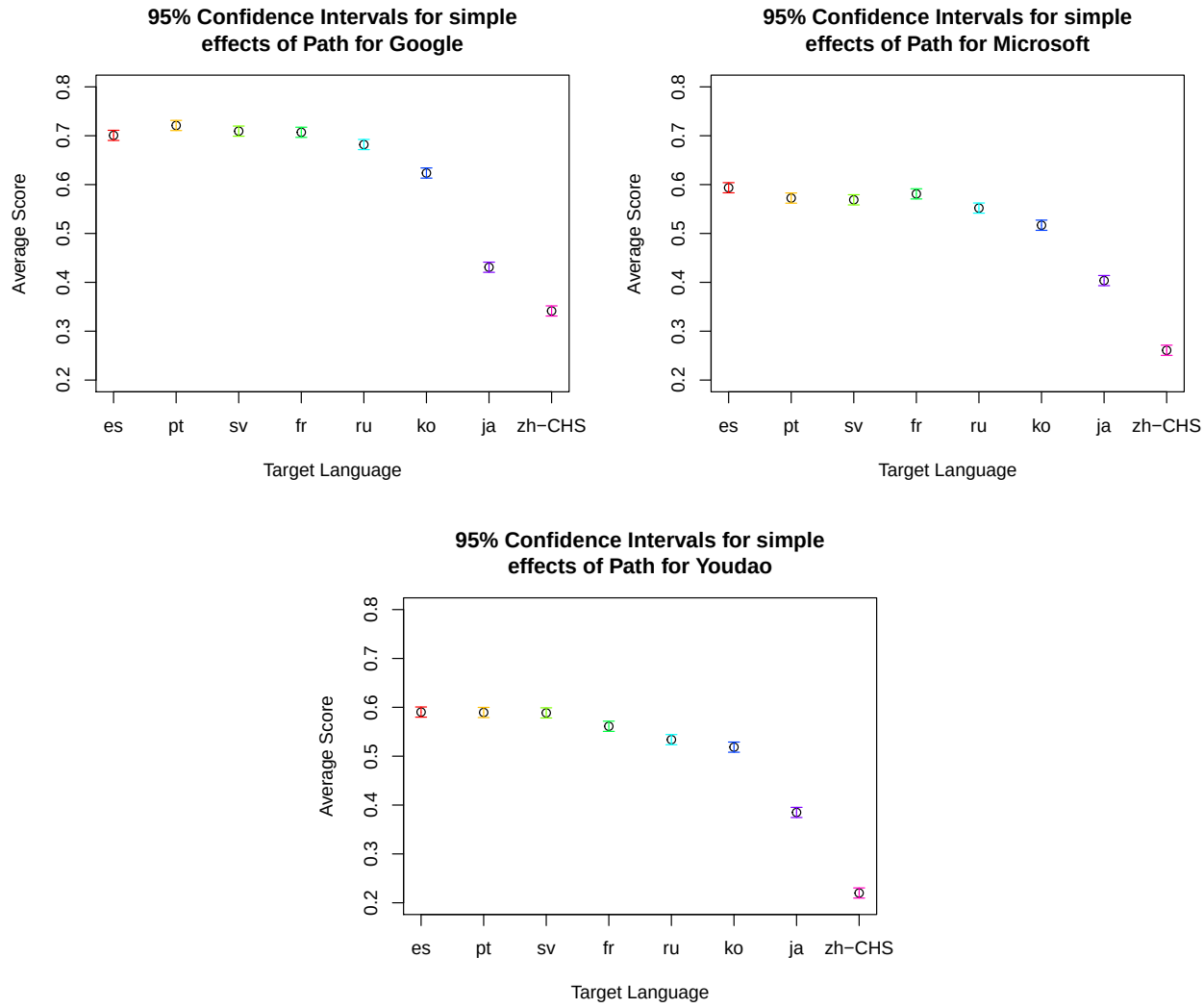

Figure 3: Simple effects of path grouped by common target language, across each translator. es: Spanish, pt: Portuguese, sv: Swedish, fr: French, ru: Russian, ko: Korean, ja: Japanese, zh-CHS: Simplified Chinese.

provide an affirmative response to both these research questions. We have proposed a new method to automatically evaluate machine translation services using a simple MR that avoids round-trip translations (RTT) [3]. The empirical results of this method consisted of 33,600 observations (involving a total of 100,800 actual machine translation outputs generated by Google, Microsoft, and Youdao translation services for one source language (English) and eight target languages). By performing statistical analysis on these results, we were able to objectively identify which services best satisfied this MR (Google Translate) and which areas faced challenges (translations into the Asian languages).

To enhance the validity of our findings, we conducted a small-scale follow-up study, which involved the use of a human assessor, one target language (Chinese), and two translation services (Google and Microsoft). It was found that the human evaluation results were consistent with the machine evaluation results reported in previous sections. On average, it took more than 30 seconds for a human assessor to evaluate one piece of translation. To perform the same task, our automated testing tool requires less than two seconds. This comparison shows that our approach is highly cost-effective.

The external validity of our findings can be enhanced by conducting larger scale experiments using different source languages and different sample sources other than Wikipedia. In future research, we will further study the correlation between the machine evaluation scores and human evaluation scores. We also plan to

[3] Having said that, a further study suggests that our results and the RTT results can often be consistent. Further discussion on the RTT results, however, is beyond the scope of this paper.

Table 3: Simple effects of path and homogeneous subsets.

	Target Language	Estimate	Group					
Google	Portuguese	0.7209871	A					
	Swedish	0.7094092	A					
	French	0.7070268	A					
	Spanish	0.7007260	A	B				
	Russian	0.6819666		B				
	Korean	0.6236494			C			
	Japanese	0.4308912				D		
	Chinese	0.3414364					E	
Microsoft	Spanish	0.5935882	A					
	French	0.5810083	A	B				
	Portuguese	0.5723823	A	B	C			
	Swedish	0.5688555		B	C			
	Russian	0.5519567			C			
	Korean	0.5168478				D		
	Japanese	0.4035918					E	
	Chinese	0.2609551						F
Youdao	Spanish	0.5902570	A					
	Portuguese	0.5894090	A					
	Swedish	0.5885621	A					
	French	0.5611871		B				
	Russian	0.5338486			C			
	Korean	0.5185094			C			
	Japanese	0.3848805				D		
	Chinese	0.2200295					E	

The standard error for each estimate is 0.005303906. Estimates not appearing in the same group are significantly different. Each of the 3 sets of 28 comparisons were made under a Tukey-adjusted family-wise error rate of 5% for each set.

Table 4: Summary of human assessed scores of 140 direct English to Chinese translations

Service	Assessed Score				Total
	0	1	2	3	
Google	0	12	26	32	70
Microsoft	0	39	23	8	70
Total	0	51	49	40	140

investigate the implications of our findings for various application areas such as cross language information retrieval and cross language sentiment analysis.

ACKNOWLEDGMENTS

This work was supported in part by a linkage grant of the Australian Research Council (project ID: LP160101691). We would also like to thank Suzhou Insight Cloud Information Technology Co., Ltd for supporting this research. We are grateful to Kenneth Russell of the University of Wollongong for his valuable comments on this work. We wish to gratefully acknowledge that the following students of the University of Wollongong contributed to part of the implementation in a preliminary study for this work: Kieran MacRae, Daniel Barnes, Shixin Wang, and Boyang Yan.

REFERENCES

[1] Milam Aiken and Mina Park. 2010. The efficacy of round-trip translation for MT evaluation. *Translation Journal* 14, 1 (2010). http://translationjournal.net/journal/51reverse.htm

[2] Earl T. Barr, Mark Harman, Phil McMinn, Muzammil Shahbaz, and Shin Yoo. 2015. The oracle problem in software testing: A survey. *IEEE Transactions on Software Engineering* 41, 5 (2015), 507–525.

[3] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit.* O'Reilly Media. http://www.nltk.org/

[4] S. G. Carmer and M. R. Swanson. 1973. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *J. Amer. Statist. Assoc.* 68, 341 (1973), 66–74. https://doi.org/10.2307/2284140

[5] T. Y. Chen, S. C. Cheung, and S. M. Yiu. 1998. *Metamorphic testing: A new approach for generating next test cases.* Technical Report HKUST-CS98-01. Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong.

[6] Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, T. H. Tse, and Zhi Quan Zhou. 2018. Metamorphic testing: A review of challenges and opportunities. *ACM Computing Surveys* 51, 1 (2018), 4:1–4:27.

[7] T. Y. Chen, T. H. Tse, and Z. Q. Zhou. 2003. Fault-based testing without the need of oracles. *Information and Software Technology* 45, 1 (2003), 1–9.

[8] David Roxbee Cox and Nancy Reid. 2000. *The Theory of the Design of Experiments.* Chapman and Hall/CRC.

[9] Anna Huang. 2008. Similarity measures for text document clustering. In *Proceedings of the 6th New Zealand Computer Science Research Student Conference (NZCSRSC 2008).* Christchurch, New Zealand, 49–56.

[10] Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings of the Workshop on Statistical Machine Translation (StatMT '06).* Association for Computational Linguistics, 102–121.

[11] Russell V. Lenth. 2016. Least-Squares Means: The R package lsmeans. *Journal of Statistical Software* 69, 1 (2016), 1–33.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 311–318. https://doi.org/10.3115/1073083.1073135
[13] Thierry Poibeau. 2017. *Machine Translation*. The MIT Press.
[14] R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[15] Sergio Segura, Gordon Fraser, Ana B. Sanchez, and Antonio Ruiz-Cortés. 2016. A survey on metamorphic testing. *IEEE Transactions on Software Engineering* 42, 9 (2016), 805–824.
[16] Tomohiro Shigenobu. 2007. Evaluation and usability of back translation for intercultural communication. In *Proceedings of the 2nd International Conference on Usability and Internationalization, Lecture Notes in Computer Science, vol 4560*. Springer-Verlag, 259–265.
[17] H. Somers. 2005. Round-trip translation: What is it good for?. In *Proceedings of the Australasian language technology workshop*. 127–133.
[18] Shengnan Zhang, Yan Hu, and Guangrong Bian. 2017. Research on string similarity algorithm based on Levenshtein distance. In *IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. 2247–2251. https://doi.org/10.1109/IAEAC.2017.8054419
[19] Zhi Quan Zhou, Shaowen Xiang, and Tsong Yueh Chen. 2016. Metamorphic testing for software quality assessment: A study of search engines. *IEEE Transactions on Software Engineering* 42, 3 (2016), 264–284.