2019

Metamorphic Relations for Data Validation: A Case Study of Translated Text Messages

Boyang Yan  
*Peng Cheng Laboratory, by932@uowmail.edu.au*

Brian Yecies  
*University of Wollongong, byecies@uow.edu.au*

Zhi Q. Zhou  
*University of Wollongong, zhiquan@uow.edu.au*

---

**Publication Details**  
Metamorphic Relations for Data Validation: A Case Study of Translated Text Messages

Abstract
In conventional metamorphic testing, metamorphic relations (MRs) are identified as necessary properties of a computer program's intended functionality, whereby violations of MRs reveal faults in the program - under the assumption that the source and follow-up inputs (test cases used in metamorphic testing) are valid. In the present study, the authors argue that MRs can also be used to validate and assess the quality of the program's input data - under the assumption that the source or follow-up inputs can be inappropriately generated. Using this new perspective, a case study in the natural language processing domain is used to explore the different types of text messages that are difficult to interpret by (Chinese-English) machine translation. A total of 46,180 short user comments on Personal Tailor (a 2013 Chinese film), collected from Douban (a popular Chinese social media platform), has been used as the primary dataset of this study, and the analysis of results demonstrates that the proposed MR-based data validation method is useful for the automatic identification of poorly translated text messages.

Disciplines
Engineering | Science and Technology Studies

Publication Details

This conference paper is available at Research Online: https://ro.uow.edu.au/eispapers1/2833
Metamorphic Relations for Data Validation: A Case Study of Translated Text Messages

Boyang Yan
Research Center of Network and Communications
Peng Cheng Laboratory
Shenzhen, China
Email: yanby@pcl.ac.cn

Brian Yecies
School of the Arts, English and Media
University of Wollongong
Wollongong, NSW 2522, Australia
Email: byecies@uow.edu.au

Zhi Quan Zhou*
School of Computing and Information Technology
University of Wollongong
Wollongong, NSW 2522, Australia
Email: zhiquan@uow.edu.au

Abstract—in conventional metamorphic testing, metamorphic relations (MRs) are identified as necessary properties of a computer program’s intended functionality, whereby violations of MRs reveal faults in the program—under the assumption that the source and follow-up inputs (test cases used in metamorphic testing) are valid. In the present study, the authors argue that MRs can also be used to validate and assess the quality of the program’s input data—under the assumption that the source or follow-up inputs can be inappropriately generated. Using this new perspective, a case study in the natural language processing domain is used to explore the different types of test messages that are difficult to interpret by (Chinese-English) machine translation. A total of 46,180 short user comments on Personal Tailor (a 2013 Chinese film), collected from Douban (a popular Chinese social media platform), has been used as the primary dataset of this study, and the analysis of results demonstrates that the proposed MR-based data validation method is useful for the automatic identification of poorly translated text messages.

Keywords: Metamorphic relation, data validation, data quality assessment, natural language processing, machine translation, sentiment analysis, oracle problem, metamorphic testing, social media, Douban.

I. Introduction

In software testing, an oracle is a mechanism against which testers can decide whether the test case execution outcomes are correct. Sometimes, an oracle is unavailable or is too expensive to be applied, which is a situation known as the oracle problem [1]. For example, software that processes big data is difficult to test due to the lack of a tangible test oracle [2], [3].

Metamorphic testing [4], [5] is a property-based software testing paradigm that is proven effective for addressing the oracle problem [1], [6], [7]. It was originally developed as a software verification method, whereby the software under test (SUT) is verified against prescribed metamorphic relations (MRs) [4], [5]. A MR is an expected relation among the inputs and outputs of multiple executions of the SUT.

Because MRs are necessary properties of the software’s intended functionality, a bug in the SUT is revealed if a MR is violated on certain test cases during testing. Metamorphic testing can be adopted by both development organizations [8], [9] and end-user programmers [10].

Xie et al. [11] found that metamorphic testing could be used not only for verification but also for validation. Zhou et al. [12] further developed metamorphic testing into a unified framework for software verification, validation, and other types of quality assessment. Metamorphic testing has been applied not only for conventional testing but also for other purposes, including cybersecurity enhancement [13], [14] and system understanding [15]. In recent years, a trend has emerged for applying metamorphic testing to machine learning and autonomous systems [16], [17], [18], [19]. In particular, Zhou and Sun combined metamorphic testing and fuzzing and detected previously unknown fatal software defects in the LiDAR obstacle-perception module of Baidu Apollo (a popular real-life self-driving platform) and reported the alarming findings eight days before Uber’s deadly crash in Tempe, AZ, USA, in March 2018 [20].

As briefly reviewed above, the mainstream metamorphic testing research community has used MRs as a type of expected property of the SUT’s intended functionality and, therefore, a violation of a MR reveals issues in the SUT. Note, however, that this treatment is under the (implicit) assumption that the metamorphic test cases being used are valid. Consider, for example, a program p(G, a, b) that finds the shortest path from an origin node a to a destination node b in an undirected graph G. A MR for p can be stated as follows: Let (G1, a1, b1) and (G2, a2, b2) be the source and follow-up inputs to p, respectively, where graph G2 is a permutation of graph G1, node a2 in G2 corresponds to node a1 in G1, and node b2 in G2 corresponds to node b1 in G1. Then p(G1, a1, b1) and p(G2, a2, b2) must return the same length for the shortest paths they found [20]. In other words, if p(G1, a1, b1) and p(G2, a2, b2) return different path lengths then p must be at fault.

In the present research, the authors observe that, in

*Corresponding author.
practical situations, it is sometimes difficult to assess the validity of the metamorphic test cases and, subsequently, the problem revealed by a MR violation could lie in the test cases (that is, the specific input to the SUT may be invalid or of poor quality) rather than in the SUT itself. For instance, if the test driver that generates isomorphic graphs is faulty then a MR violation in the above shortest-path example could have been caused by $G_2$ not being isomorphic to $G_1$ (rather than $p$ being faulty). This observation suggests that MRs could be used to validate or assess the quality of data instead of verifying software correctness, which can be useful for a variety of fields.

To evaluate the usefulness of the above new perspective of MRs, a case study is conducted in the natural language processing domain, exploring the different types of text messages that are difficult to interpret by machine translation.

The rest of this paper is organized as follows: Section II introduces the background of this research. Section III describes the metamorphic relation identified for data validation and assessment in the natural language processing domain. Section IV explains the procedure of empirical evaluation, and the results are presented in Section V. Finally, Section VI concludes the paper and discusses future research directions.

II. BACKGROUND

A. The Application Domain and Dataset of the Case Study

With the increasing global spread of Web 2.0 platforms and applications—some containing internal translation functions—social and digital media users have encountered language and cultural barriers during cross-language communication. Simply put, there is now a blue ocean of user-generated content (UGC) in various languages that awaits instantaneous translation for the average user. Notwithstanding, it remains impossible to manually translate the massive volume of UGC that is generated on a day-to-day basis. Yet, a host of machine translation tools are currently available for achieving such ends, such as Google Translate, Microsoft Translator, Yandex Translate, Baidu Translate and Youdao Translate, etc. As the creation and spread of UGC in different languages has increased, so too has the number and popularity of machine translation services that ultimately could facilitate global integration. The economic and cultural benefits of this transformation are high, and machine translation systems have already been shown to increase the significance of international trade [22]. Much previous research focuses on an economic context, ranking the quantity of machine translation services in some fields [23]. However, to date, only limited research has been conducted on the evaluation of difficult text translations (which is the theme of the present research).

In this research, the original (Chinese) dataset is sourced from Douban (Chinese: 豆瓣) (https://www.douban.com), China’s largest and most popular interest-oriented social media platform of its kind, featuring shared UGC on movies, books, music and other forms of popular culture. A total of 46,180 short Chinese-written comments (collected through the Douban public APIs in July 2018) linked to Personal Tailor (Chinese: 私人订制), a 2013 Chinese film, are used as the primary dataset to investigate the translation quality of different translation tools. This particular movie is selected for the wide-ranging colloquial language that it has inspired among Douban’s user responses. The short comments posted and shared on Douban by registered users (and read by the public) are up to a maximum of 140 Chinese characters.

According to a survey [24], at the time of Douban’s launch in 2005, only about 7% of adults in the United States were using social media; however, after the advent of Android devices and Apple’s iPhone App Store, social media usage rapidly exploded. As of 2015, 65% of adults in the United States were using social media. After China’s homegrown tech firms Baidu, Alibaba and Tencent (collectively known as BAT) became Web 2.0 titans in their own right, Chinese users of China’s online and mobile platforms soon followed this upward trajectory. Naturally, UGC on each of the different social media platforms contains unique language (word level) and stylistic characteristics, such as colloquial vocabulary that enables convenience of expression. Researchers in the fields of digital humanities and computer science have an urgent need to pursue in-depth studies in this rapidly changing arena, which may open possibilities for new theoretical approaches, methods and understandings.

In the present study, the authors focus on three major machine translation services in the world: Google Translate (U.S., https://translate.google.com), Yandex Translate (Russia, https://translate.yandex.com) and Baidu Translate (China, https://fanyi.baidu.com), assessing the particularities of translating Chinese (the no. 1 most spoken language in the world) to English (the no. 2 most spoken language in the world) and the difficulties that arise from doing so.

Evaluating the advantages and limitations of machine translation services poses a number of difficulties that usually require input from language experts who are familiar with the multiple languages of the corpus. However, language expertise alone also involves human emotional judgments, thereby making objective assessment of human language somewhat difficult, not to mention automated assessment without a test oracle [1]. While recent research suggests that metamorphic testing can be effective for addressing some of the difficulties in automatic testing of machine translation services [23, 25], the present research uses a MR from a different perspective as explained previously, aiming to develop a method of automatically identifying poorly translated text messages without relying on a language expert. This is an important step for researchers and stakeholders in the authors’ respective disciplines.
B. Sentiment Analysis

The aim of text sentiment analysis is to determine the attitudes of writers with respect to particular topics and the overall contextual polarity or emotional reaction. It usually involves the use of natural language processing and machine learning techniques to determine the possibility of positive or negative opinions [26]. Sentiment analysis is useful in an e-business context. For example, business managers can analyze customers’ attitudes, as to whether they like or dislike their products or service. Also, government can use sentiment analysis to obtain citizen perspectives.

III. Metamorphic Relation for Validation and Assessment of Data

The present research employs a metamorphic relation of sentiment analysis, and then uses this MR, together with a sentiment analysis tool, to evaluate the quality of machine-translated text messages. The MR is described as follows:

**MR1**: Let $L_1$ and $L_2$ be two different natural languages. Let $s_1$ and $s_2$ be text messages written in languages $L_1$ and $L_2$, respectively. Let $A$ be a sentiment analysis tool that can work for both languages $L_1$ and $L_2$. If $s_2$ is an accurate translation of $s_1$ then the sentiment analysis results produced by $A$ for $s_1$ and $s_2$ should be similar.

**Remark**: $MR_1$ is a metamorphic relation for $A$. Different from conventional metamorphic testing, however, the intention of designing this MR is not to test $A$, but rather to evaluate (to a certain extent) the semantic (dis)similarity between $s_1$ and $s_2$: If $A$ is a high quality sentiment analysis tool then violation of $MR_1$ could indicate that $s_2$ is a poor translation of $s_1$. In this way, $MR_1$ is used for the validation and quality assessment of data (namely, whether $s_2$ is a good translation of $s_1$).

IV. Evaluation Procedure

Three distinct machine translation services (namely, Google, Yandex and Baidu) are used to translate the original dataset (denoted by $Chinese_{\text{original}}$, which is the set of 46,180 user comments gathered from Douban) from Chinese into English, thus generating three additional datasets, each containing 46,180 translated comments in English. These additional datasets are named $English_{\text{Google}}$, $English_{\text{Yandex}}$ and $English_{\text{Baidu}}$.

Using the Google sentiment analysis API (https://cloud.google.com/natural-language/docs/sentiment-tutorial), which works for both the Chinese and English languages, each and every comment in the above four datasets are analyzed against $MR_1$, whereby the elements of $MR_1$ are mapped as follows:

$L_1$: Chinese; $L_2$: English; $s_1$: a comment in $Chinese_{\text{original}}$; $s_2$: a translated comment in $English_{\text{Google}}$, $English_{\text{Yandex}}$ or $English_{\text{Baidu}}$; $A$: the Google sentiment analysis API.

Thus, a total of $46,180 \times 4 = 184,720$ sentiment analyses have been performed: 46,180 analyses for the original Chinese comments and 138,540 analyses for their English versions translated by the three machine translation services. These 184,720 results are then analyzed against $MR_1$ (after noise reduction to eliminate meaningless results).

The Google sentiment analysis API produces two separate values: score and magnitude. The score values range between $-1$ and $1$ (a larger value corresponds to a more positive attitude). In this research, $MR_1$ is considered violated if one side of the score (that is, $A(s_1)$ or $A(s_2)$ in $MR_1$) is greater than $0.7$ and the other side is smaller than $-0.7$. The large threshold value ($\pm 0.7$) is chosen in order to identify the (Chinese comment, English translation) pairs whose sentiment analysis results differ greatly. This choice is somewhat arbitrary, and a different tester could choose a different threshold value or range as they consider appropriate for their specific needs. It should also be noted that limited scope of this research prevents an analysis of magnitude, which distinguishes neutral comments.

V. Evaluation Results

The Venn diagram of Fig. 1 shows the distributions of $MR_1$ violations. A total of 296 pairs of translations have been identified (Google: 137; Baidu: 134; Yandex: 129). An investigation into these results shows that the majority of them reveal erroneous translations (although some sentiment analysis issues are also discovered), which means that the adopted approach is effective. The three machine translation services are owned and operated by the largest tech firms of their kind in their respective countries, and as expected they cater to different international users and audiences. Notwithstanding, all three services contain a number of unique as well as shared lexical and grammatical shortcomings. The following aspects reveal some of the most common problems shared by all of them.
A. Vocabulary Errors

Some of the biggest types of mistakes are due to the abuse and/or incorrect usage of vocabulary, chaotic sequences and incomplete phrases. Examples include:

1) Fragmented sentences:

The original Chinese movie comment: “2013最后一场电影就为恶心画一句号吧” which means: “This is [my] last movie in 2013. Let’s put a complete stop on nausea,” The Google translation is: “Let’s take a look at the last movie of 2013.”

As one can see, the original Chinese comment is meant to be extremely negative in nature. The hidden meaning is that this movie is nauseating. The Google translated sentence is fragmented and incomplete, translating only the first part of the original sentence.

2) Misunderstanding of vocabulary:

The original Chinese movie comment: “這片子真垃圾就最后一点看起来比较有意义” which means: “This is really a crappy movie [or: This movie is real garbage]; just the last part [of the movie] is quite innovative.” The Google translation is: “This piece of real garbage on the last point looks more innovative.” And, the Yandex translation is: “This film is really garbage; it finally looks relatively new.” Finally, the Baidu translation is: “The film is really garbage and it looks new at the last point.”

The front part of the original Chinese sentence is negative, and the second half of the sentence is more positive. As such, these three machine translation tools each generate their own vocabulary misunderstandings, which can be depicted as:

- “movie” → “piece”
- “the last part [of the movie]” → “finally” or “the last point.”

Here is another example:

The original Chinese movie comment: “剧本不好，梗比较生硬，有点无厘头” which means: “The scenario is bad, the punchline is stiff, a bit making no sense.” The Google translation is: “The script is not good, terrier stiffer, a bit nonsense.”

And the Yandex translation is: “The script is bad, stem relatively stiff, a bit does not make sense.”

Finally, the Baidu translation is: “The script is not good, the stem is relatively stiff, a little silly.”

The original Chinese sentence is clearly negative. All three tools mistranslate the “punchline” (in Chinese: 梗) because this particular Chinese character is Internet lingo, a.k.a. slang. In addition, all three translation tools frequently misinterpret a range of professional and literary terms throughout the corpus, such as “春晚” (the Spring Festival Gala Show). Suffice it to say that each tool needs to improve its engagement with slang and Internet lingo.

B. Grammatical Errors

In addition to generating fragmented sentences partly from misunderstanding vocabulary, the machine translation tools reproduce grammatical mistakes, such as sequence disorder and improper location of adjectives and adverbs. This creates confusion among subjects, objects, predicates and other attributive concepts.

1) Illogical connections:

The original Chinese movie comment: “明明是搞笑片最后弄成了教育片” which means: “Obviously it is a comedy but it’s been made into an educational film in the end.” The Baidu translation is: “It’s a funny movie that finally makes an education film”—here, the word “that” appears as an illogical connection between the subject and its subordinate clause.

2) Incomplete ingredients:

The original Chinese movie comment: “中间确实无聊得很，但是结尾不赖” which means: “It’s really boring in the middle, but it’s not bad at the end.” The Yandex translation is: “The middle of really boring very, however at the end.”

3) Sequence disorder:

The original Chinese movie comment: “内涵挺有点的就是拍的太肤浅” which means: “Quite a bit meaningful but made too superficial.” The Google translation API fails to translate this sentence. The Yandex translation is: “The connotation quite a bit is shoot too superficial for” and the Baidu translation is: “The connotation is quite a little that is it is too superficial.”

These poor translation results are beyond the simple mishandling of particular professional or industry terms that might cause some translation confusion. These translations contain incomplete sentence composition and confusing word order. As a result, machine translation has yet to replicate colloquial and proper language results faithfully—at least when applied to the popular style of communicating with short comments.

One obvious explanation for these documented differences between human and machine language is the large gap between colloquial language characters shared in Douban film review comments and the insufficient (but expanding) vocabulary library accessed during each tech firm’s machine translation processes. Here, there is a paradox between the size of the text and the abilities of the platforms. While the online translation of text comments and reviews is relatively short, the appearance of colloquial conventions poses challenges for each tool to achieve more accurate translations. Furthermore, although the sentence structure of these short comments seems simple, the existence of compound sentences and their rich and complex patterns
remains hidden. Today, there are major differences in thinking between Chinese and English language vocabulary and sentence structures, which is difficult for intelligent machines to discern, but this is changing as machine tools are continuously improved. For the moment, however, the shortcomings discussed herein are among the main difficulties facing the quality of machine translation and related validation activities.

VI. Conclusion and Future Work

The contributions of this paper to the software engineering literature are summarized as follows:

Metamorphic relations (MRs) have been used conventionally to test computer programs in the context of metamorphic testing, under the assumption that the input data (test cases) are valid whereas the program under test could be faulty. In the present research, the authors have developed a new perspective whereby MRs can also be used to validate and assess the quality of the input data, under the assumption that the input data could be invalid or of poor quality and that the computer programs (for which the MRs are identified) are of high quality.

A case study has been conducted to evaluate the usefulness of the proposed concept and perspective. The application domain is in natural language processing, whereby a MR has been designed for a sentiment analysis tool. With this MR, the authors’ intention is not to test the sentiment analysis tool, but rather to validate and assess the quality of the input data passed to the sentiment analysis tool. These data are pairs of text messages where each pair has the form (original Chinese sentence, translated English sentence). The empirical results show that this MR-based data validation strategy is useful, as it has effectively identified poorly translated sentences.

In future research, more theoretical and empirical studies need to be conducted for MR-based data validation in a variety of application domains.

The contributions of this paper to the machine translation literature are summarized as follows:

In general, Google, Yandex and Baidu translation tools can translate the basic meanings of a sentence. They are relatively fluent and understandable. However, the present study shows that all three of these machine translation tools share some poor qualities, including the overlooking of individual words, improper grammar selection, and failure to account for particular language expression in natural speaking. Simply put, they can fail to recognize the characteristics of colloquial text. All of these factors contribute to mistranslation, making it extremely difficult (at this time of writing in the emergent AI-era) to handle aptly the special meanings of certain common words in a particular text. Henceforth, the colloquial language contained in Chinese social media movie comments may require different translation methods in order to perfect the translation of their contents and meanings.

As with previous work on metamorphic testing for machine translation (MT4MT) [23], [25], the MR-based data validation and evaluation method developed in this research is useful for detecting translation errors without needing to depend on a language expert.

It should be noted, however, that MT4MT is focused on the verification and validation of the translation software rather than the input data, although test data are generated during this process. In contrast, the focus of the present research is not on the testing of the translation software, but rather is on the evaluation of data (the translated text messages in the case study). In other words, this research is focused on validation approaches for a given dataset rather than test case generation techniques for a given program. The proposed data validation methodology, therefore, does not need to involve the concepts of “software under test” or “test cases generation.”

This paper shows that, to a certain degree, the proposed methodology can account for errors in machine translated sentences and also the so-called paradox of polarity, a scenario in which the actual polarity of results involving both language datasets conflicts with the expected outcomes.

Due to differences in the ways of thinking and expressing between the Chinese and English languages, as well as cultural differences and limitations of translation datasets, researchers in the fields of digital humanities and computer science need to continue investigating machine translation methods for mutual understanding. Translation tools are expected to demonstrate a series of language problems into the near-future. Thus, machine translation researchers must look beyond their narrow disciplines and seek new opportunities to work with fellow linguistic researchers. Together, this cross-disciplinary approach has great potential for investigating different datasets, and conducting lexical and grammatical analysis on a globally expanding dynamic corpus of social media and UGC.

ACKNOWLEDGMENTS

This work is supported in part by the project “PCL Future Regional Network Facilities for Large-Scale Experiments and Applications” (PCL2018KP001), and also supported in part by a linkage grant of the Australian Research Council (Project ID: LP160101691).

REFERENCES


