Difficult Text Analysis of Machine Translation Services base on Metamorphic Testing in Social Media

1st Boyang Yan
Research Center of Network and Communications
Peng Cheng Laboratory
Shenzhen, China
yanby@pcl.ac.cn

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

***Abstract*—A huge amount of text comments are posted on different topics in Social Media every day. These topics are discussed in different languages by different language speakers. Most people encounter language and culture barriers when engaging in cross-language communication. Cross-language machine translation is useful for global integration. However, most of people are chosen human translation services. The reason is human translation services are accuracy and reliable compare with machine translation. Most research only focuses on ranking the quantity of machine translation services but little research has conducted on difficult translation text evaluation. This research explores what kind of text is difficult to translate for machine translation services base on movie comments data. It is useful for improve the quality of machine translation services to fill this research gap. This research is based on the Metamorphic Testing method to establish a testing model which using machine translation service translate original test dataset from one language to another language. After, using sentiment analysis tool to analyize original test dataset and translated dataset, the results should be same polarization (positive or negative). If the results are opposite polarization, that means this sentence is difficult for machine translation. As a result, people will able to use this testing model finding difficult sentences and doing specific optimiztion.**

***Index Terms*—Metamorphic Testing, machine translation, sentiment analysis, machine translation quantity testing, evaluation of machine translation services difficult**

I. Introduction

Machine translation services has been becoming more and more widely used, also more and more popular. Most people encounter language and cultural barriers during cross-language communication. There are lots of text and documents on different languages need to translation every day. It would be impossible to translation the huge amount of data generated manually. Nowadays, there are lots of machine translation tools are available in the world, such as Google translation, Bing translation, Yandex, Baidu translation and Youdao translation and so on. In this research, only compare and evaluation Google translation, Yandex translation and Baidu translation difficults. Those three translation tools are typical and the most widespread to used. Google translation tool come from American, Yandex come from Russia, and Baidu come from China. Accounting to Pesu said machine translation tools can product better results on European languages compare with Asian language [1]. So, Chinese to English translation tool is the main kind of translation languages to analysis translation difficults in the paper. Evaluation of machine translation services difficults usually need language expert, who need well-known both languages, to participate. However, language expert also involves human emotional judgment. Automatic assessment human language is naturally difficult because of without a test oracle [2]. In this paper, achieving a testing modle to automatic assessment without language expert. Metamorphic testing(MT) is one of property-based software quality testing method, which already be appoved effective for addressing the non-oracle problem, such as testing the quality of search engine and the quality of Unmanned Aerial Vehicle(UAV) flight control application and so on. Therefore, decideing metamorphic testing to find machine translation services difficults in non-oracle sitation. And more specifically, this research raise two questions.

- Q1: What is current sitation of the quantity of Chinese to English machine translation?
- Q2: What is current machine translation difficults between Chinese and English?

The rest of paper is organized as three parts. Firstly, domonstration the quantity of Chinese to English translation services, which are Google translation, Yandex translation and Baidu translation. This part addresses Q1. secondly, describtion testing model about finding the difficults of machine translation. Thirdly, analyzes the experimental results and discussion. This part will answer Q2.

II. Background

A. Metamorphic Testing

Metamorphosis Testing(MT) a method for generating test cases, as well as test results verification. The most importance component is the metamorphic relation (MR). MR is the target application's necessary properties of function in relation to multiple inputs and their expected outputs. When people want to assess the the accuracy of $\sin$ function. For example, $\sin(2.7)$ is very difficult to make a correctness judgment from mathematics aspect. If using Metamorphic Testing method to testing $\sin$ function will reduce computational costs and more efficient. The testing procedures are:

1) set a Metamorphic Relation: such as $\sin(\alpha) = \cos(\frac{\pi}{2} - \alpha)$
2) $\sin(2.7)$ and $\cos(\frac{\pi}{2} - 2.7$ should have same output, if the outputs are different. We can say, the failure have been detected.

However, when using II-A Metamorphic Testing method set a metamorphic relation, which is will less cost . If make a correctness judgment

which are expected relations among the inputs and outputs of multiple executions of the intended program's functionality.

are an approach for generating both test cases and test results. A central element is a set of metamorphosis relationships, which are necessary properties of the objective function or the algorithm in relation to multiple inputs and their expected outputs.

Since its first publication, we have witnessed a rapidly increasing body of work examining metamorphic testing from various perspectives, including metamorphic relation identification, test case generation, integration with other software engineering techniques, and the validation and evaluation of software systems.

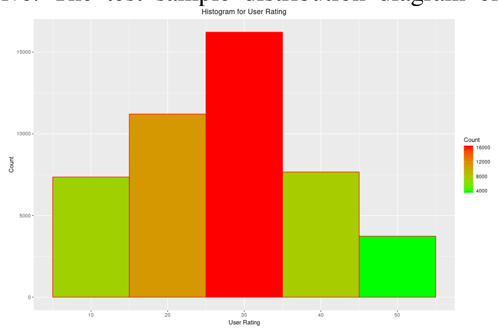
In software testing, an inability to determine if software is behaving correctly, or producing the correct output, is called the oracle problem [1]. Metamorphic testing (MT) is an approach that can alleviate the oracle problem [6, 10], MT has been investigated and adopted by a growing number of researchers and practitioners [7, 16, 20, 21, 28], successfully uncovering software problems, even in extensively tested systems [8, 9, 19]. Central to MT is a set of metamorphic relations (MRs), which are expected relations among the inputs and outputs of multiple executions of the intended program's functionality. Instead of examining the behaviour or output for an individual input, MT checks the SUT against selected MRs, with violations of an MR indicating the presence of a fault. An example MR for a database management system is that the system should return the same results for a query with the search condition "A and B" and a query with the search condition "B and A".

B. Sentiment Analysis

III. Domonstration the current quantity of Chinese to English translation services

A. Test Sample

All of test sample came from Douban, one of biggest social networking service platforms in China. This social website attracts more than one hundred million active visitors per month, and has amassed over sixty-five million registered users. We then employ the Douban public Application Programming Interfaces (APIs) to access Chinses-written comments. A typical data structure of harvested comment is shown as a tuple: [Rating, Raw comments]. Totally, comments have got 46180 in the corpus. User rating total have 5 groups, which are 10, 20, 30, 40 and 50, from negative to positive. The test sample distribution diagram on below.



As you can see, the majority of comments allocate on rating 30. In addition, there have got more negative comments compare with positive comments.

B. Testing procedures

1) using three of machine translation services to translate Chinese original movie comments to English translated movie comments.

$$P_{(OriginData)} \rightarrow P'_{(GoogleTranslation)}$$

$$P_{(OriginData)} \rightarrow P'_{(YandexTranslation)}$$

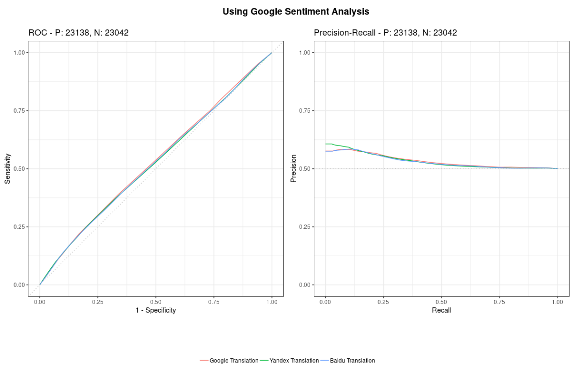
$$P_{(OriginData)} \rightarrow P'_{(BaiduTranslation)}$$

2) Using Google sentiment analysis tool to analysis $P_{(OriginData)}$, $P'_{(GoogleTranslation)}$, $P'_{(YandexTranslation)}$ and $P'_{(BaiduTranslation)}$. Google sentiment analysis APIs will get 2 values, which are Score and Manitude. The range of Score is between -1 and 1. If Score more close to 1 means this movie comment more positive, as well as, if Score more close to -1 means this movie comment more negative.
3) Using user rating values (10, 20, 30, 40 or 50) to check Google Sentiment Analysis(SA) results $P_{(OriginData)}$ is True or False. For example, user rating = 10 and Google Chinese SA score between -1 and -0.6 (mean True) user ranking = 10 and Google Chinese SA score bigger than -0.6 (mean False). The decideing True table on below. Else Not in True table all False.

user rating	Google SA score	True/False
10	$[-1, -0.4]$	True
20	$[-0.8, 0]$	True
30	$[-0.4, 0.4]$	True
40	$[0, 0.8]$	True
50	$[0.4, 1]$	True

4) Using those True or False values as vector combining with Google English SA scores (based on $P'_{(GoogleTranslation)}$), Google English SA scores (based on $P'_{(YandexTranslation)}$) and Google English SA scores (based on $P'_{(BaiduTranslation)}$) draw 3 Receiver operating characteristic (ROC) graphics and 3 Precision-recall curves (PRC) graphics.

   ROC curve is often used in evaluation the clinical performance of a biochemical test. The ROC curve is based on a series of different binary classifier with the true positive rate (sensitivity) as the Y-axis and the false positive rate (1-specificity) as the X-axis [3]. The traditional evaluation must be divided into two categories, and then statistical analysis is performed. The ROC curve is different from the traditional evaluation method. Instead, an intermediate state is allowed. The test results can be divided into multiple ordered classifications then statistically analyzed. However, visual interpretation and comparisons of ROC curves based on imbalanced data sets can be misleading. An alternative to a ROC curve is a precision-recall curve (PRC). PRC might be a better choice for imbalanced datasets [4].



   This graphic show those three of machine translation tools all achieve poor translation results.

5) Calculate Area Under The Curve (AUC) values for ROC and PRC.

TABLE I
GOOGLE TRANSLATION

Curve Types	AUCs
ROC	0.5307797
PRC	0.5328503

   The AUC is between 1.0 and 0.5. The better diagnostic effect will be close to 1.

TABLE II
YANDEX TRANSLATION

Curve Types	AUCs
ROC	0.5251734
PRC	0.5322736

TABLE III
BAIDU TRANSLATION

Curve Types	AUCs
ROC	0.5258386
PRC	0.5302736

- AUC has lower accuracy from 0.5 to 0.7
- AUC has a certain accuracy from 0.7 to 0.9
- AUC has higher accuracy at above 0.9

When AUC=0.5, it means that the diagnostic method is completely ineffective and has no diagnostic value [5].

- For ROC AUCS: It shows **Google Translation tool** better than **Baidu Translation tool** better than **Yandex Translation tool**
- For PRC AUCS: It shows **Googl Translation tool** better than **Yandex Translation tool** better than **Baidu Translation tool**
- The number of true value: 23042
- The number of false value: 23138

Alought the dataset is looks balanced, ROC diagram can be trusted. The ranking of machine translation services' quantity are NOT reliable. The reason is three of translation services have lower accuracy. In another word, working not properly correct.

C. *finding the difficults of machine translation*

D. *Testing procedures*

1) using three of machine translation services to translate Chinese original movie comments to English translated movie comments.

$$P_{(OriginData)} \rightarrow P'_{(GoogleTranslation)}$$

$$P_{(OriginData)} \rightarrow P'_{(YandexTranslation)}$$

$$P_{(OriginData)} \rightarrow P'_{(BaiduTranslation)}$$

2) Using Google sentiment analysis tool to analysis $P_{(OriginData)}$, $P'_{(GoogleTranslation)}$, $P'_{(YandexTranslation)}$ and $P'_{(BaiduTranslation)}$. The result are $SA_{(OriginData)}$, $SA'_{(GoogleTranslation)}$, $SA'_{(YandexTranslation)}$ and $SA'_{(BaiduTranslation)}$
3) This three of relations should NOT be opposite attitude
   - R1:

$$SA_{(OriginData)} \longleftrightarrow SA'_{(GoogleTranslation)}$$

   - R2:

$$SA_{(OriginData)} \longleftrightarrow SA'_{(YandexTranslation)}$$

- R3:

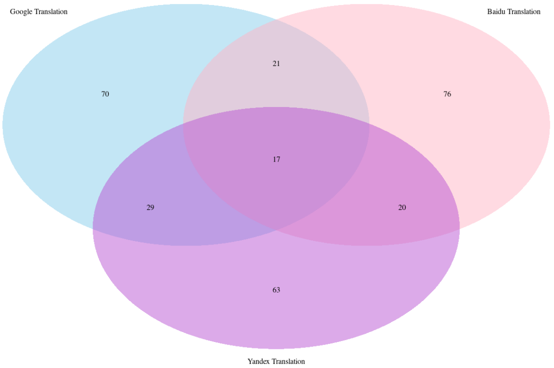
$$SA_{(OriginData)} \longleftrightarrow SA'_{(BaiduTranslation)}$$

E. Analysis

R1, R2 and R3 have got failures decideing by one side greater than 0.7 and another side smaller than -0.7. In this paper, using veen diagram for show failures distribution.

TABLE IV
TRANSLATION FAILURES DISTRIBUTION

Types	Number Of Failures
google Translation	137
Yandex	129
Baidu	134
$Google \cap Baidu$	38
$Google \cap Yandex$	46
$Yandex \cap Baidu$	37
$Yandex \cap Baidu \cap Google$	17



REFERENCES

[1] D. Pesu, Z. Q. Zhou, J. Zhen, and D. Towey, “A monte carlo method for metamorphic testing of machine translation services,” in *Proceedings of the 3rd International Workshop on Metamorphic Testing*. ACM, 2018, pp. 38–45.

[2] Z. Zhou, S. Xiang, and T. Y. Chen, “Metamorphic testing for software quality assessment: A study of search engines.” *IEEE Trans. Software Eng.*, vol. 42, no. 3, pp. 260–280, 2016.

[3] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (roc) curve.” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[4] J. Davis and M. Goadrich, “The relationship between precision-recall and roc curves,” in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

[5] Baidu. receiver operating characteristic curve. [Online]. Available: https://baike.baidu.com/item/ROC