The paradox of Machine Translation base on Sentiment Analysis polarity: using Metamorphic Testing

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. It is necessary for people to use machine translation tools in global integration. However, accuracy and reliable of machine translation tools services is paying close attention by users. The quantity of machine translation services and difficult translation text evaluation have been a hot area of research for several years. It is important for improving the quality of machine translation services to narrow the research gap. The purpose of the paper is to explore what kind of text is difficult to translate for machine translation services base on movie comments data in Douban Movie Community Website. Metamorphic Testing is main method for establishing a testing model which using Google Sentiment Analysis tool to analyze original Chinese movie comments dataset and English translated datasets, which are Google Translation dataset, Yandex Translation dataset and Baibu Translation dataset, both side (Chinese and English) should be same emotional 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. The research result shows that the three translation tools overall quality are poor based on Receiver operating characteristic(ROC) curve and precision-recall curve (PRC). In addition, there are outstanding language problems in machine translation, computer science researchers must be combined with linguistic researchers, so further research of Machine translation should pay enough attention to linguistics research, the expansion of corpus, vocabulary differentiation, grammatical analysis and so on, should be deeply and comprehensively studied.

Index Terms—Metamorphic Testing, machine translation, MT4MT, sentiment analysis, machine translation quantity testing, evaluation of machine translation services difficult, natural language, Receiver operating characteristic(ROC) curve, precision-recall curve (PRC)

I. INTRODUCTION

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. Machine translation services has been becoming more and more widely used, also more and more popular. It is necessary for global integration to use Cross-language machine translation tools. In especial, machine translation system has significantly increased international trade [1]. Most research only focuses on ranking the quantity of machine translation services in some field [2]–[6] but little research has conducted on difficult translation text evaluation. In this paper, all of testing data (Origin Chinese dataset) from DOUBAN, which is an Social Media website about books, movies, music, to study the translation quality of different translation tools. Douban film is China's largest and most authoritative film sharing and comment community [7]. Social Media has been becoming more and more widely used. Accounting to Perrin's survey, there are only $7 social media in 2005. However, social media usage increase rapidly, the$ Media website comment text from the language word level has its unique style, such as the text has colloquial vocabulary for the convenience of expression. At present, it is urgent to carry out in-depth study on this field, which may open up the new field of theoretical research. In this study, we only fouced on three of Machine Translation tools, which are Google, Yandex and Baidu. Google translation tool come from American, Yandex come from Russia, and Baidu

come from China. This three tools all come from the top three powerful country. Accounting to Pesu said machine translation tools can product better results on European languages compare with Asian language [2]. So, form 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 [8]. 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 alreadly be appoved effective for addressing the 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. 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) is a method for generating test cases, as well as test results verification [9]. The most importance component is the metamorphic relation (MR) [10]. MR is the target application's necessary properties of function in relation to multiple inputs and their expected outputs. MT has been researched through more and more researchers constantly strive toward and adopted by industries and organizations such as Adobe, NASA and the National Institute of Standards and Technology [11]. In software testing research field, an incapacity to decide, software product the correct output, is called the oracle problem [12]. This usually means cannot provide exact correctness reference data. such as, machine translation. Huge and complexity systems, does not have reference data for proving function's correctness, is very common. 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. There is a testing procedures' example for $\sin$ function.

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, this MR have been break, maybe the failure have been detected.

However, when using II-A's testing procedures. Someone maybe ask, $\cos$ function also not reliable, how can use a unreliable function to test another function's correctness. There have the boundedness. However, both function have got failures at same time, that is small probability event.

B. Sentiment Analysis

Sentiment analysis is a part of text data mining. The aim of sentiment analysis is to determine the attitude of speakers or writers with respect to particular topics or the overall contextual polarity or emotional reaction to a text document. It is usually equated with opinion mining, which involves the use of natural language processing and machine learning to ascertain the possibility of positive or negative opinions [13]. Sentiment analysis is useful for analyzing a huge amount of data relating to personal opinions. It can be used in an e-business context. For example, business managers can analysis customers' attitudes, as to whether they like or dislike their product or service. Also, government can use sentiment analysis to analyze citizen perspectives. In this paper, we will using Google Sentiment Analysis tool to prove the the quantity of Machine Translation. The details of Testing Modle will talk in III-B .

III. DOMONSTRATION THE CURRENT QUANTITY OF CHINESE TO ENGLISH TRANSLATION SERVICES

A. Test Sample

All of test sample came from Douban[1], 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 [14]. 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

One the below, will explain more details for figure 2.
1) using three of machine translation services to translate Chinese original movie comments to English translated movie comments.

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

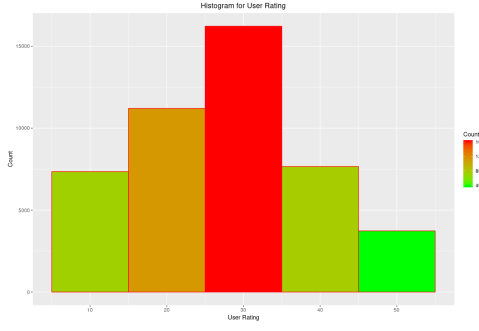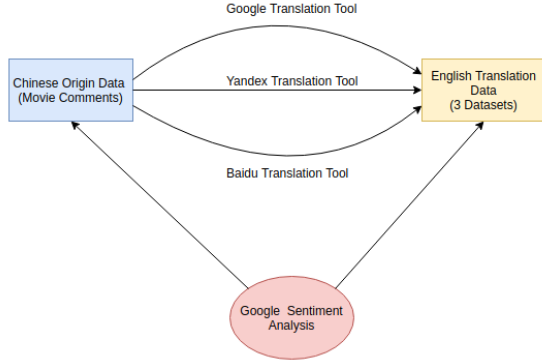[1] Douban Website: https://www.douban.com/

Fig. 1. User Rating Histogram

Fig. 2. Testing Procedures

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

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

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. In this paper, we have not analysis Manitude, which for distinction mix and neutral.

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.

As you can see, the Google SA score ranges have got some overlap because overlap can decrease the results, machine translation evaluation correctness, influence by the accuracy of Google Sentiment Analysis tool.

4) Using those True or False values as vector combining with Google English SA scores (based on $P'_{(GoogleTranslation)}$), Google English SA scores (based

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
10	$(-0.4, 1]$	False
20	$[-1, -0.8)$	False
20	$(-0.8, 1]$	False
30	$[-1, -0.4)$	False
30	$(0.4, 1]$	False
40	$[-1, 0)$	False
40	$(0.8, 1]$	False
50	$[-1, 0.4)$	False

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 [15]. 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 [16]. 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 [17]. This graphic show those three of machine translation tools all achieve poor translation results.
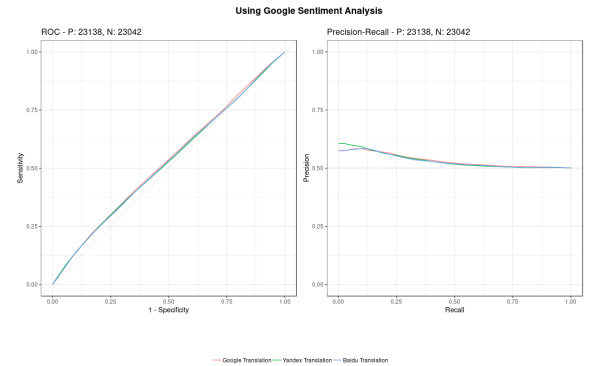

Fig. 3. ROC and PRC graphics

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

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

TABLE II GOOGLE TRANSLATION	
Curve Types	AUCs
ROC	0.5307797
PRC	0.5328503

TABLE II
GOOGLE TRANSLATION

Curve Types	AUCs
ROC	0.5307797
PRC	0.5328503

TABLE III
YANDEX TRANSLATION

Curve Types	AUCs
ROC	0.5251734
PRC	0.5322736

TABLE IV
BAIDU TRANSLATION

Curve Types	AUCs
ROC	0.5258386
PRC	0.5302736

TABLE V
ROC AND PRC ACCURACY JUDGMENT

AUC value	Accuracy
$[0.5, 0.7]$	lower accuracy
$[0.7, 0.9]$	certain accuracy
$[0.9, 1]$	higher accuracy

show on Table: 5

When AUC=0.5, it means that the diagnostic method is completely ineffective and has no diagnostic value [18].

C. Current Translation Tool Results Analysis

In the testing Procedures 3, which is the judgment of Google Chinese Sentiment Analysis(SA) results is True or False, there are totally have find 23042 of true value, as well as 23138 of false value. 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. However, there are still have the ranking of machine translation services' quantity on the below.

1) For ROC AUCS: **It shows Google Translation tool better than Baidu Translation tool better than Yandex Translation tool**

2) For PRC AUCS: **It shows Googl Translation tool better than Yandex Translation tool better than Baidu Translation tool**

IV. FINDING THE DIFFICULTS OF MACHINE TRANSLATION

A. Testing procedures

The first two step is same with, showing the current quantity of Machine Translation Tools in Section III. Which are get translated datasets and sentiment analysis for 4 datasets(Chinese Origin dataset, Yandex translation dataset, Google translation dataset and Baidu translation dataset). The only different is filting all of opposite polarization (very positive or very negative) datasets and create three of Metamorphic Relation (MR).

1) MR1:

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

2) MR2:

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

3) MR3:

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

B. Analysis

MR1, MR2 and MR3 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 VI
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

V. LANGUAGE ANALYSIS

Google Translation, Yandex Translation and Baibu Translation are created by different countries for different audiences, that is to say, the users of groups are different. But Google Translation, Yandex Translation and Baidu Translation exist lexical and grammatical problems, and some mistranslation is almost coincide. The following aspects are the main common problems.

A. Vocabulary Errors

The points of mistake are mainly manifested as the abuse of vocabulary, incorrect usage, the chaotic sequence, missing

vocabulary,as well as improper selection of words. Examples:

1) fragmention sentence:

- Chinese Origin Movie comment is:
    "2013最后一场电影就为恶心画一句号吧"
- The <u>correct of the translation</u> is:
    "This is the last movie in 2013. Let's make a full stop for all of nauseating movie"
- The <u>Google Translation</u> is:
    "Let's take a look at the last movie of 2013"

This is extremely negative comment. The hidded meaning is, this is nauseating movie. The Google translated sentence is fragmented and incomplete only translated the first part of original sentence.

2) misunderstand of vocabulary:

- Chinese Origin Movie comment is:
    "这片子真垃圾就最后一点看起来比较有新意"
- The <u>correct of the translation</u> is:
    "This is really crap movie. However, the last part of movie is considerable innovative.
- The <u>Google Translation</u> is:
    "This piece of real garbage on the last point looks more innovative"
- The <u>Yandex Translation</u> is:
    "This film is really garbage it finally looks relatively new"
- The <u>Baidu Translation</u> is:
    "The film is really garbage and it looks new at the last point."

The original Chinese sentence front part is negative, but the back half sentence is positive. This three of machine translation tools have got lots of vocabulary misunderstand. such as,

$$movie \rightarrow piece$$

$$last\ part\ of\ movie \rightarrow finally\ or\ the\ last\ point$$

Besides,

- Chinese Origin Movie comment is:
    "剧本不好，梗比较生硬，有点无厘头。"
- The <u>correct of the translation</u> is:
    "The scenario is not good, the punchline of movie is stiff and a little bit does not make sense."
- The <u>Google Translation</u> is:
    "The script is not good, terrier stiffer, a bit nonsense."
- The <u>Yandex Translation</u> is:
    "The script is bad, stem relatively stiff, a bit does not make sense."
- The <u>Baidu Translation</u> is:
    "The script is not good, the stem is relatively stiff, a little silly."

The original Chinese sentence is clear negative. All of three tools have got problem with punchline of movie, "梗" is the internet lingo word. All of tools need to improve the Internet lingo words. In addition, three of translation tools also have got others misinterpretation of professional terms and literary terms. such as "春晚" (Spring Festival Gala Show).

B. Grammatical Errors

The machine translation have got lots of grammatical mistake, such as sequence disorder, improper location of adverbial attributive. The subject, the predicate, the attributive, the adverbial, and the object is incomplete, etc.

1) Illogical connect:

- Chinese Origin Movie comment is:
    "明明是搞笑片最后弄成了教育片"
- The <u>correct of the translation</u> is:
    "Obviously it is a comedy,but it made into educational films finally"
- The <u>Baidu Translation</u> is:
    "It's a funny movie that finally makes an education film"

The "that" is illogical to connect Subject part and subordinate clause.

2) Incomplete ingredients: Origin Chinese sentence is "中间确实无聊得很，但是结尾不赖". The original sentence meaning "It's really boring in the middle part, but it's not bad at the end". The Yandex Translation "The middle of really boring very, however at the end".

3) Sequence Disorder: "想过是烂片，只是没想到能烂到这种地步，太应付了，一个星期拍完的吧" Google "Think of a bad movie, but did not expect to be able to rot to this point, too coping, a week finished it" Yandex "Thought is a bad film, just didn't think it could be rotten to the point where too cope with, a week finished." baidu "That is a bad film, but did not expect to come to this rotten, too for a week, finished."

The problem of sentence structure.Incomplete sentence composition, confusion of words order.

For example, original sentence meaning"A few small cups of grape juice, it is not very fresh".GoogleTranslation "A few small cups of grape juice, very fresh"in completely different meaning;

Take another example,original sentence meaning " The first story is good, but the other is too contrived". YandexTranslation"On the first story good the other is too contrived";

Besides,For example,original sentence meaning "The first half was good with heavy irony, but the last apology was a little redundant".

BaiduTranslation"The first half was good - the irony was very heavy, and the last apology was a little redundant". Improper location of clauses"-". The misuse of conjunctions"and".

The problem of sentence structure.Causality is unclear. For example,original sentence meaning "I can not ignorant conscience to give too few stars,because I am a gentle person".GoogleTranslation "I am a gentle person, and I really do not have the heart to give too few stars. . ."

Take another example,original sentence meaning "Bad,I will never see Feng Xiaogang's movie again." YandexTranslation "Sucks, feel Feng Xiaogang movie later never to see."

从以上以豆瓣电影评论为分析对象来讨论这三个翻译软件的译文共同的问题，从整体上看，无论是用词还是语法都一定的错误导致许多译文意思不准确，内容残缺，结构错误。

From the overall points of view, above the Douban film review as the analysis object to discuss GoogleTranslation, YandexTranslation and BaiduTranslation tools common problems, whether the usage of words or grammar are a certain error caused by many translation inaccurate meaning, incomplete content, and the wrong structure.

The reasons are colloquial language characters of Douban film review text and insufficient machine vocabulary.Although the length of the movie commentary sentences is short, the vocabulary is not difficult and the sentences'structure seems simple, there have got lots of colloquial speech and many of words are not formal, which is the main reason for the low quality of translation results. Lots of compound sentence is hidded. It is more difficult for intelligent machines translation services.Although the length of the movie commentary sentences is short, the vocabulary is not difficult and the sentences'structure seems simple, there have got lots of colloquial speech and many of words are not formal, which is the main reason for the low quality of translation results. Lots of compound sentence is hidded. It is more difficult for intelligent machines translation services.

Although the length of the movie commentary sentences is short, the vocabulary is not difficult and the sentences' structure seems simple, there have got lots of colloquial speech and many of words are not formal, which is the main reason for the low quality of translation results. Lots of compound sentence is hidded. It is more difficult for intelligent machines translation services.

Although the online translation text comments and review is short in length, but more colloquial is serious, the use of words is not standard, so vocabulary translation is more difficult.Although the sentence structure seems simple, but hides the compound sentence, but the sentence pattern is rich and complex.Besides,Differences in thinking between Chinese and English is more difficult for intelligent machines, which is the main difficulty in the quality difficulties of translation

VI. CONCLUSION AND LIMITATIONS

In general, Google Translation, Yandex Translation and Baidu Translation tools can still translate the basic meaning of a sentence. They are basically fluent and understandable. However, all of three Machine Translation tools are poor quality and similar. As you can seen, this tesing model is possible to detected Machine Translation failures without language expert. But, there have got limitation in the accuracy of Sentiment Analysis tool, which difficult to distinguish minimal change of emotion. As a results, this testing model will missing most of Machine Translation failures, particularly, neutral emotional sentences. This tesing model is good for decideing strong emotion sentences translation correctness. Accounding to failures of machine translated sentence. The main problem is all of three translation services have get the individual words are not properly selected and the language expression does not conform to the natural speaking. They do not recognize the characteristics of colloquial text. Some of the mistranslations are almost overlap. So it is not possible to correctly handle the special meaning of certain common words in a particular text. And the colloquial language of movie comments should adopt different translation methods in order to perfect the content of the translation. It is still necessary to strengthen the processing of sentence structure and translation expression.

VII. FUTURE WORK

Due to the differences way of thinking between Chinese and English languages. In addition, cultural differences and the limitation of translation corpus. Machine translation tools have serious language problems. Machine translation researchers must be work with linguistic researchers, who should be paid more attention to the corpus, lexical analysis and grammatical analysis.

REFERENCES

[1] E. Brynjolfsson, X. Hui, and M. Liu, "Does machine translation affect international trade? evidence from a large digital platform," National Bureau of Economic Research, Tech. Rep., 2018.

[2] 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.

[3] M. Aiken and M. Park, "The efficacy of round-trip translation for mt evaluation," *Translation Journal*, vol. 14, no. 1, 2010.

[4] P. Koehn and C. Monz, "Manual and automatic evaluation of machine translation between european languages," in *Proceedings of the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2006, pp. 102–121.

[5] H. Somers, "Round-trip translation: What is it good for?" in *Proceedings of the Australasian Language Technology Workshop 2005*, 2005, pp. 127–133.

[6] T. Shigenobu, "Evaluation and usability of back translation for intercultural communication," in *International Conference on Usability and Internationalization*. Springer, 2007, pp. 259–265.

[7] 杨桦, "豆瓣阅读自出版平台优势浅析," 新媒体研究, vol. 24, pp. 44–45, 2018.

[8] 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.

[9] Z. Zhou, D. Towey, P.-L. Poon, and T. Tse, "Introduction to the special issue on test oracles," 2017.

[10] T. Y. Chen, T. H. Tse, and Z. Q. Zhou, "Fault-based testing without the need of oracles," *Information and Software Technology*, vol. 45, no. 1, pp. 1–9, 2003.

[11] Z. Q. Zhou and L. Sun, "Metamorphic testing for machine translations: Mt4mt," in *2018 25th Australasian Software Engineering Conference (ASWEC)*. IEEE, 2018, pp. 96–100.

[12] J. Brown, Z. Q. Zhou, and Y.-W. Chow, "Metamorphic testing of navigation software: A pilot study with google maps," in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[13] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, p. 25, 2017.

[14] siteworthtraffic, "Douban.com traffic worth," *http://www.siteworthtraffic.com/report/douban.com*, Accessed 19 April 2017.

[15] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine." *Clinical chemistry*, vol. 39, no. 4, pp. 561–577, 1993.

[16] 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.

[17] 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.

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