Metamorphic Testing in Cross-language Sentiment Analysis for Social Media

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

XXXXXXXXX
XXXXXXXXXX
Springfield, USA
Email: XXXXXXX

***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 opinion mining is useful for global integration. However, most research only focuses on English language sentiment analysis, but little research has been conducted on sentiment analysis in languages other than English. This research explores using machine translation and sentiment analysis tools to fill this gap. The research identifies a combination of tools which will enable people to understand different language speakers' attitudes (positive or negative), emotions and opinions. This research is based on the Metamorphic Testing method to establish a testing model for finding which machine translator service combined with which English sentiment analysis service can obtain reliable sentiment analysis results for non-English speakers who do not have sentiment analysis tools to analysis their own language. As a result, people will able to use Machine Translation and English Sentiment Analysis to conduct big data analysis in multi-language Social Media.**

1. Introduction

Social Media has been becoming more and more widely used. There are lots of text comments on different discussion topics every day. It would be impossible to analyses the huge amount of data generated manually. These topics are discussed by speakers of different languages, from different cultural backgrounds, further complicating any analysis. Most people encounter language and cultural barriers during cross-language communication. In this research, the use of machine translation and sentiment analysis tools to solve this problem of analyzing cross-cultural and cross-language data is explored and discussed. 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 [1]. 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 a word, sentiment analysis is coming into widespread use. As Dr. Haiyun mentions, English language sentiment analysis research has undergone major developments in recent years [2]. However, less research has been undertaken in other languages, such as Chinese. Today, a lot of English language sentiment analysis theories have been developed. Also, a variety of Machine Translation tools is available, such as Google translation, Bing translation and Yandex translation. Machine Translation uses computational linguistic programs and natural language processing theory [3]. However, nobody working in a combination of these two fields of research has undertaken non-English sentiment analysis. Therefore, the research described and discussed in this paper aims to create a testing model to find the best-combination of English sentiment analysis tools and machine translation tools to obtain reliable sentiment analysis results from non-English texts, for non-English speakers who do not have such sentiment analysis tools to analyze their own language. Eventually, everyone will be able to understand different language speakers' attitudes (positive or negative), emotions and opinions. The purpose of the study have two aspects. The first aims of this research is to compare and analysis Google, Youdao, Baidu, Bing and Yandex translation tools, which one is the best machine translation tool. Second aim is creating a testing model to find the best-combination of English sentiment analysis tools and machine translation tools to obtain reliable sentiment analysis results. Third aim is to create our own sentiment analysis model for recognizing English data belonging to positive, negative, neutral or mix classification.

2. Critical Review of the Literature

This research consists of three components; measuring machine translation service quality; testing sentiment analysis service quality; finding the best compound mode for machine translation service and sentiment analysis service. As a result, this section will focus on a review of literature

review about machine translation, testing methodology and sentiment analysis.

2.1. Testing in Machine translation

there are two research articles about testing modeling of machine translation (MT).

2.1.1. Round Trip Translation method. As Somers argues, an Around Trip Translation (RTT) method has been establish to detect the quality of machine translation [4]; for example, testing English to Chinese translation tools. Firstly, an English to Chinese translation tool is use to translate test data to Chinese. It is then used to translate Chinese data back to English. Finally, compare the similarity for two English data set. They also mention two metrics of similarity, BLEU and F-score, to judge the translations. The limitations of RTT model are it cannot distinguish the best MT tool from a group of poor MT tools as well as it cannot find which sentences are easier for translation and which sentences are harder for translation.

2.1.2. A Monte Carlo Method for machine translation services. Another article is about using third-party language to test the quality of machine translation [5], for example, if testing an English to Chinese translation tool. Firstly, random choose an Intermediate third-party language. Secondly, translation English test data to the third-party language, after translating, the third-party language to Chinese. These two steps constitute one path. Another path is translation from English directly to Chinese. In the end, the two path results need to compare similarity. In this article, the main finding is that Google Translate is the best machine translation compare with Yandex, Youdao as well as Bing. In addition, the better results to be produced in European languages compare with Asian languages, use ANOVA Statistics method and Pairwise T tests giving this conclusion. In my experiment, I also got Google Translator is the best machine translation compare with Yandex and Baidu. Pairwise T tests also can be useful for finding best-combination of English sentiment analysis tools and machine translation tools to obtain reliable sentiment analysis results. The highlight of this model is using third-party language, which can decide preference language of machine translation.

3. Testing methodology

Accounting to Sethi, there are two categorized testing techniques, which are Static Testing and Dynamic Testing. The Dynamic Testing are divided into three categories, which are Functional Testing, Structural testing and Non-Functional Testing [6]. In my research project, I will focus on Functional Testing on this project.

3.0.1. Metamorphic Testing. a)This research project is based on Metamorphic Testing. Metamorphic Testing is for testing function correctness. A research article written by Zhou in 2016 clearly explains what Metamorphic Testing is. As Zhou explains, Metamorphic Testing (MT) is a property-based software testing method developed for automated test case generation and automated results verification, based on the effects of some expected properties of the target program [7]. These properties, recognized as metamorphic relations (MRs), serve as essential relations between the inputs and outcomes of multiple executions of the target program. For instance, calculator function correctness will be established if the input is 1 + 1, and the result is 2. In this example people can easily make the judgment, whether the calculator is functioning correctly or not. However, people will not be able to easily make this judgment if the input is sin (3.7). In a generally acknowledged, Sin (3.7) = sin (3.7 + 360) is correct. In metamorphic testing, the name of 3.7 is the test case. The name of 3.7 + 360 is the follow-up test case. Metamorphic Relation is the relationship between two input test cases as well as the two outputs. Metamorphic testing is based on Metamorphic Relation. The two outputs need an existing mathematical relation. In this example, the relation is "=". However, MR does not must be an equation, it also can be a relation. The advantage of Metamorphic testing (MT) method; addressing the test oracle problem; testing case generation problem. The disadvantage is that it cannot detect memory leak or some others insensitivity failure situation. However, Metamorphic Testing is appropriate for testing translation tools and sentiment analysis tools.

3.0.2. Effectiveness of Metamorphic Relations. [8] there is another article which was written by Zhou about Effectiveness of Metamorphic Relations in 2013. The main purpose of reading this article is trying to find which Metamorphic Relations can be the most efficient detecting failures. Round Trip Translation and a Monte Carlo Method can be seen as two Metamorphic Relations. This article is based on white box testing, which have source code, as well as the most important conclusion is if the Metamorphic Relations can get bigger distance (dissimilarity) that will have more chance to detect failures. In other words, MRs with very different initial and follow-up execution are more likely to detect failures than those with similar initial and follow-up executions. The concept of "difference" are defined in namely coverage Manhattan distance (CMD), frequency Manhattan distance (FMD), and frequency Hamming distance (FHD) in regard to adaptive random testing (ART), where CMD metric on the basis of branch coverage execution profiles performs the best fault detection effectiveness. The advantage of this article is suitable for finding the most effectiveness of Metamorphic Relations in White-box. However, this article is not suitable for Black Box Testing. The reason is Black-Box Testing have not source code available, so it cannot calculate the program's distance. In this research project, translation tools have NOT source code available, this article is not suitable for this research accordingly.

3.0.3. White-box VS Black-box. There is another research article written by Henard in 2016. Talking about the difference between black-box testing and white-box testing.

Henard (2016) have done some research for difference between white box testing and black - box testing in 2016. They have two finding is useful in my research black-box testing and white-box testing performance just have a little difference (at most 4 fault detection rate difference). They also found black box testing and white-box testing the overlap is very high. The first 10 of the prioritized test data already agree on at least 60 of the faults found. As the result, this research article has given me a lot of ideas of how the similarity between white box testing and black-box testing. I still have opportunity for compare those three modeling's, which one is better [9].

4. Objectives and Scope

The aim for this research is to achieve a method to find out the combination between machine translation tools and English sentiment analysis model can obtain the result, which is the most reliable and efficient, for those non-English speakers, to fill the blank and gap of lacking of cross-language sentiment analysis tool. The main aim can be divided into 4 sub-aims. I. Advertisement Detection Detect advertisements and junk contents among mass of data from social media texts. Removing unimportant data can both reduce the amount of the size of whole dataset to save processing time, and get rid of contents that is of no use to our sentiment analysis, which can be also regarded as noise data. II. Data Preprocessing and Feature Extraction Preprocessing data is to segment texts into words and select those words which is helpful and sensitive in sentiment analysis. E.g. keywords, important punctuation marks, emotion symbols. Also, normalization operations will be taken to convert the keywords into it root form, which can reduce the size of lexicons of models to a large extend. Then, extracting features of those data being preprocessed as inputs for sentiment analysis model built by us. III. Sentiment Analysis Modelling Develop a model for sentiment analysis which includes a lexicon placing emphasis on social media texts and a machine learning model for analyzing sentiment. The lexicon should consider about the main feature of social media texts below: short-text styled, sparsity of contents and concluding emoticons. This can make our model performs better than the other ones which focus on general texts. For the machine learning model, we aim to design a model which considers about efficiency, accuracy and reliability. IV. Cross-language Translation and Model Testing For our source of data coming from social media which is in different languages, finding a better performed tool for translation is of vital importance. We aim to find the best performed tool for each respective combination of languages, so that we can have closer meaning according to the origin language. With the final text data, testing should be designed to test the real performance of the model we build. And based on the results of testing, we can have optimization on relevant domains of our research.

5. Methodology and Procedure of the Study

d)Getting test data: getting movie reviews data from social media website for finding the best compound mode for machine translation service and sentiment analysis services. e)Testing Machine translation services quality and Sentiment Analysis services quality — In this part focus on testing Yandex, Baidu, Google Machine Translation tools, as well as, Baidu, Google sentiment analysis tools. This is testing model, which is according to Metamorphic Testing Method. When we testing Machine Translation, we need to assume sentiment analysis tools perfectly correct. We can compare correlation coefficient between both sides of sentiment analysis results for getting which machine translation are better. There is an example for testing Chinese to English machine translation. I. Using Google, Baidu, Yandex translation tools, translated original Chinese data to English data. II. Using same sentiment analysis tool analysis original Chinese dataset and translated dataset. III. Calculate correlation coefficient between Chinese sentiment analysis results and English sentiment analysis results. IV. Compare correlation coefficient values. If value is bigger than others, we can say this translation tool, which use in original dataset to English dataset, can achieve better results than others. For testing sentiment analysis tools, we can using same model. We need assuming Machine Translation tools are prefect correct. If right side of sentiment analysis result is opposite attitude with left side of sentiment analysis result. We can say we are detected one failure. f)Finding the best compound mode for machine translation service and sentiment analysis services In this part, we totally have 6 kinds of compound model, which are Google translation with Google sentiment analysis; Yandex translation with Google sentiment analysis; Baidu translation with Google sentiment analysis; Google translation with Baidu sentiment analysis; Yandex translation with Baidu sentiment analysis and Baidu translation with Baidu sentiment analysis. We can using mean-square error (MSE) and Receiver operating characteristic (ROC) compare with user rate get the best compound model. g)Create own sentiment analysis model Accounting to Liu said, creating sentiment analysis model have five steps. I will basis on this five steps for create my own model [10]. i) Get Terms - Reduce review to the list of keywords ii) Filtering - Remove unnecessary keywords that will not add value for sentiment analysis, such as is, but, it etc. iii) Find the Base Word - Convert all inflections to their root word iv) Make Features - Use the root words as features to indicate the positiveness or negativeness v) Classifier - Train a classifier to predict positivity.

6. Implications and Significance of the Problem

Social Media has been becoming more and more widely used. Accounting to Perrin's survey, there are only 7we also find most of people are interest in different language speakers' opinions and attitudes. In addition, most peo-

ple encounter language and cultural barriers during cross-language communication. As Dr. Haiyun mentions, English language sentiment analysis research has undergone major developments in recent years [2]. However, less research has been undertaken in other languages, such as Chinese. Today, a lot of English language sentiment analysis theories have been developed. Also, a variety of Machine Translation tools is available, such as Google translation, Bing translation and Yandex translation. Machine Translation uses computational linguistic programs and natural language processing theory [3]. However, nobody working in a combination of these two fields of research has undertaken non-English sentiment analysis. Therefore, the research described and discussed in this paper aims to create a testing model to find the best-combination of English sentiment analysis tools and machine translation tools to obtain reliable sentiment analysis results from non-English texts, for non-English speakers who do not have such sentiment analysis tools to analyze their own language. Eventually, everyone will be able to understand different language speakers' attitudes (positive or negative), emotions and opinions.

7. Expected Outcomes

Testing Model: giving the result about which machine translation tool can achieve more accurate translated result. I guess Google ¿ Yandex ¿ Baidu, my model also can finding which of these sentence are hardly for machine translation. Maybe, all of machine translation tools will change the human being's attitude (positive and passive attitudes trend to neutral). For movie review part, User can give impartial rate for movie review, because my testing model for finding the best compound mode for machine translation service and sentiment analysis services dependent on user rate. If user rates are impartial. I can get credible conclusion for which one is the best compound mode. I can use same compound mode to analysis others data. My testing model can widely use in e-business aspect contributed to globalization.

8. Test Data

Total have 46180 movies reviews.

Ranking	Number of Test Data	Percentage
Ranking 10	7353	15.92 %
Ramking 20	11209	24.27 %
Ranking 30	16223	35.13 %
Ranking 40	7663	16.59 %
Ranking 50	3732	8.08 %

9. IQR, Mean, Median, Q1, Q3, lower extreme, upper extreme, mean slope, median slope

9.1. Baidu Chinese Sentiment analysis Positive Probability Base On (Chinese Origin Data)

Ranking	IQR	Mean	Median	Q1	
10	0.3426533	0.2404781	0.1789910	0.0489987	0.3916
20	0.3817485	0.2949725	0.2489580	0.0878135	0.4695
30	0.4504405	0.3980354	0.3808120	0.1516095	0.6020
40	0.5255830	0.5123376	0.5385910	0.2461685	0.7717
50	0.5368080	0.5712128	0.6188395	0.3101650	0.8469

- Mean Slope

0.008788344

- Median Slope

0.0116933

9.2. Baidu Sentiment analysis Positive Probability Base On (Google Translated Data)

ranking	IQR	Mean	Median	Q1	Q
10	0.0992950	0.5109406	0.512129	0.4495650	0.548860
20	0.1083630	0.5229484	0.518486	0.4562780	0.564641
30	0.1311445	0.5396073	0.528391	0.4674360	0.598580
40	0.1795885	0.5673052	0.543921	0.4839135	0.663502
50	0.1873947	0.5882316	0.557420	0.5017473	0.689142

- Mean Slope

0.001989388

- Median Slope

0.00116017

9.3. Baidu Sentiment analysis Positive Probability Base On (Yandex Translated Data)

- Mean Slope

0.002034071

- Median Slope

0.00139647

ranking	IQR	Mean	Median	Q1	Q
10	0.1095430	0.5180638	0.515600	0.4527060	0.562249
20	0.1132460	0.5348823	0.524710	0.4697750	0.583021
30	0.1437290	0.5512503	0.532877	0.4769425	0.620671
40	0.1861070	0.5825192	0.555963	0.4933645	0.679471
50	0.1965927	0.5959489	0.569797	0.5031990	0.699791

9.4. Baidu Sentiment analysis Positive Probability Base On (Baidu Translated Data)

- Mean Slope

0.002024859

- Median Slope

0.00135257

ranking	IQR	Mean	Median	Q1	Q
10	0.0998390	0.5242141	0.5189600	0.4636380	0.56347
20	0.1116160	0.5328661	0.5240450	0.4693390	0.58095
30	0.1476325	0.5502528	0.5343010	0.4763575	0.62399
40	0.1813430	0.5819689	0.5544210	0.4933320	0.67467
50	0.1975785	0.6009056	0.5714005	0.5055505	0.70312

9.5. Google Chinese Sentiment Analysis Score Base On (origin Data)

- Mean Slope

0.01635146

- Median Slope

0.016

ranking	IQR	Mean	Median	Q1	Q3	lowerExtreme	upperExtreme
10	0.7	-0.2344349	-0.2	-0.6	0.1	-1.65	1.25
20	0.7	-0.1092783	0.0	-0.5	0.2	-1.55	1.25
30	0.7	0.1289959	0.1	-0.2	0.5	-1.25	1.55
40	0.7	0.3244682	0.4	0.0	0.7	-1.05	1.75
50	0.8	0.3662647	0.4	0.0	0.8	-1.2	2.00

9.6. Google English Sentiment Analysis Score Base On (Google Translated Data)

- Mean Slope 0.0153968
- Median Slope 0.015

ranking	IQR	Mean	Median	Q1	Q3	lowerExtreme	upperExtreme
10	0.7	-0.33609411	-0.4	-0.7	0.0	-1.75	1.05
20	0.6	-0.23656883	-0.2	-0.6	0.0	-1.5	0.90
30	0.6	-0.05040375	0.0	-0.4	0.2	-1.3	1.10
40	0.6	0.15573535	0.1	-0.1	0.5	-1.00	1.40
50	0.6	0.23759378	0.2	0.0	0.6	-0.90	1.50

9.7. Google English Sentiment Analysis Score Base On (Yandex Translated Data)

- Mean Slope 0.01467965
- Median Slope 0.015

ranking	IQR	Mean	Median	Q1	Q3	lowerExtreme	upperExtreme
10	0.7	-0.33472052	-0.4	-0.7	0.0	-1.75	1.05
20	0.6	-0.22842359	-0.2	-0.6	0.0	-1.5	0.90
30	0.6	-0.05227147	0.0	-0.4	0.2	-1.3	1.10
40	0.6	0.14357301	0.1	-0.1	0.5	-1.00	1.40
50	0.6	0.21326367	0.2	0.0	0.6	-0.90	1.50

9.8. Google English Sentiment Analysis Score Base On (Baidu Translated Data)

- Mean Slope

0.01418969

- Median Slope

0.012

ranking	IQR	Mean	Median	Q1	Q3	lowerExtreme	upperExtreme
10	0.7	-0.28185775	-0.3	-0.7	0.0	-1.75	1.05
20	0.5	-0.18345972	-0.1	-0.5	0.0	-1.25	0.75
30	0.6	-0.01048511	0.0	-0.3	0.3	-1.20	1.20
40	0.6	0.17350907	0.1	-0.1	0.5	-1.00	1.40
50	0.6	0.24914255	0.2	0.0	0.6	-0.90	1.50

10. Assessing Machine translation tool quality

10.1. Method

1) Compare correlation coefficient between Chinese sentiment analysis results and English sentiment analysis results by each
   a) Using Google, Baidu, Yandex translation tools, translated original Chinese data to English data
   b) Using same sentiment analysis tool analysis original chinese dataset and translated dataset
   c) Calculate correlation coefficient between Chinese sentiment analysis results and English sentiment analysis results
   d) Compare correlation coefficient values. if value is bigger than others, we can say this translation tool, which use in original dataset to English dataset, can achieve better results than others.

10.1.1. Result.

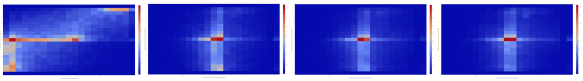
- Base on Google sentiment analysis tool

	Google Score for Google translated d
Gooogle Score for origin data	0.512 (Pearson Correlations) p-value:
Google Score for origin data	0.381 (Kendall Correlations) p-value:
Google Score for origin data	0.504 (Spearman Correlations) p-value
Gooogle Score for origin data	0.512 (Point Biserial) p-value: 0.0

- Google translation tool quality ¿ Yandex translation tool quality ¿ Baidu translation tool quality
- Base on Baidu sentiment analysis tool

	Baidu Positive Probability
Baidu Positive Probability for origin data	0.288 (Pearson Correlation
Baidu Positive Probability for origin data	0.188 (Kendall Correlation
Baidu Positive Probability for origin data	0.271 (Spearman Correlati
Baidu Positive Probability for origin data	0.288 (Point Biserial) p-va

- Google translation tool quality ¿ Yandex translation tool quality ¿ Baidu translation tool quality



11. Conclusion

The conclusion goes here.

Acknowledgments

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.