Metamorphic Testing in Cross-language Sentiment Analysis for Social Media

Boyang Yan
School of Computing and Information Technology
University of Wollongong
Wollongong, Australia
by932@uowmail.edu.au

Xiaoxia Pu
School of Computing and Information Technology
University of Wollongong
Wollongong, Australia
xp816@uowmail.edu.au

Xudong Zhang
School of Computing and Information Technology
University of Wollongong
Wollongong, Australia
xz944@uowmail.edu.au

Helene Tran
School of Computing and Information Technology
University of Wollongong
Wollongong, Australia
ht185@uowmail.edu.au

***Abstract*—Huge amounts of text comments are posted on different topics in Social Media everday. 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 conbination 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.**

***Index Terms*—sentiment analysis, machine translation, Metamorphic Testing, Social Media, Cross-Language, Cross-Culture**

I. 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 enounter language and cultureal barriers during cross-language communication. In this paper, the use of machine translation and sentiment analysis tools to solve this problem of analysing 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 analysise 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. Manchine 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.

II. Part A: Annotated bibliography

III. Literature Review

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.

A. Testing in Machine translation

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

a) 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 distingicush 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.

b) 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.

B. 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.

a) Metamorphic Testing: 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.

b) 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.

c) White-box VS Black-box: [9] There are 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 modelings, which one is better.

IV. Research Proposal [1]

A. TITLE

Metamorphic Testing in Cross-language Sentiment Analysis for Social Media

B. Background and Research Problems

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 enounter language and cultureal barriers during cross-language communication. In this research, the use of machine translation and sentiment analysis tools to solve this problem of analysing 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 analysise 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. Manchine 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

[1]This research proposal is based on Boyang Yan's Master of Research application (research proposal section) in 2017

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. It is building by VADER, NLTK, LIWC, ANEW or the General Inquirer.

C. Critical Review of Literature

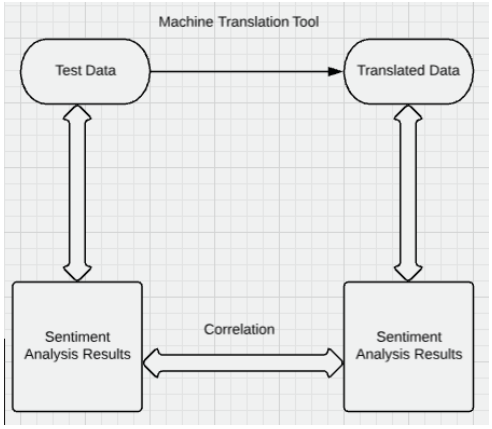
D. Aims and Objectives

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.

1) 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.
2) 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.
3) 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.
4) 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.

E. The procedure of the study

1) Survey
   The research conducts a survey to different languages backgrounds people, about what they thinking about globalization, does they want to know different cultural backgrounds people opinions and attitudes. Does they enounter language and cultureal barriers during cross-language communication. The purpose of survey is Machine Translation tools and Sentiment analysis Tools are useful for people daily life and established connection between different cultural people communication.
2) Getting test data
   if in the survey, we can get most of people enounter language and cultureal barriers and insterest in others culture, we can start getting movie reviews data from social media website for finding the best compound mode for machine translation service and sentiment analysis services
3) Testing Machine translation services quality and Sentiment Analysis services quality
   In this part fouce 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 assuming sentiment analysis tools are profect correct. We can compare correlation coefficient between both side of sentiment analysis results for getting which machine translation are better. There is an example for testing Chinese to English machine translation.

   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.

For testing sentiment analysis tools, we can using same model. We need assuming Machine Translation tools are profect 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.

4) 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.
5) 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].
   a) GetTerms - Reduce review to the list of keywords
   b) Filtering - Remove unnecessary keywords that will not add value for sentiment analysis, such as is, but, it etc
   c) Find the Base Word - Convert all inflections to their root word
   d) Make Features - Use the root words as features to indicate the positiveness or negativeness
   e) Classifier - Train a classifier to predict positivity

a) Implications or significance of the problem: Social Media has been becoming more and more widely used. Accounting to Perrin's survey, there are only 7% American adults are use social media in 2005. However, social media usage increase rapily, there are 65% American adults are use social media untill 2015 [11]. In our survey, we also find most of people are insterest in different language speakers' opinions and attitudes. In addition, Most people enounter language and cultureal 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. Manchine 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.

b) Expected Outcomes:

V. Survey - Questionnaire

Link : https://www.surveymonkey.com/r/GHBDPBL

a) : The following questionnaire on Survey Monkey composed of 10 questions is used for our research work :

1) What is your gender ?
   a) Male
   b) Female
   c) Other

   Initial Code : Gender

2) What are your areas of interest ? (many answers possible)
   a) Sport
   b) Music
   c) Watching Videos / Films
   d) Reading
   e) Cooking
   f) Doing Shopping
   g) Chatting
   h) Playing Video Games
   i) Other (please specify)

   Initial Code : Interest areas

3) How often do you use social media ?
   a) Always
   b) Usually
   c) Sometimes
   d) Rarely
   e) Never

   Initial Code : Frequency of social media use

4) Which social media do you use regularly ? (many answers possible)
   a) Facebook
   b) Twitter
   c) YouTube
   d) Instagram
   e) WeChat
   f) LinkedIn
   g) Weibo
   h) I don't use social media
   i) Other (please specify)

   Initial Code : Types of social media

5) Which language(s) can you read ?
   a) English
   b) French
   c) Chinese

d) Spanish
e) Vietnamese
f) Arabic
g) Indian
h) Russian
i) German
j) Other (please specify)

Initial Code : Language skills

6) Are you interested in other cultures ?
   a) Extremely interested
   b) Very interested
   c) Somewhat interested
   d) Not so interested
   e) Not at all interested

   Initial Code : Attitude towards culture

7) Are you willing to try to understand a comment in an unknown language ?
   a) Very likely
   b) Likely
   c) Neither likely nor likely
   d) Unlikely
   e) Very unlikely

   Initial Code : Behavior towards unknown languages

8) How often do you read others' opinion in social media posts ?
   a) Always
   b) Usually
   c) Sometimes
   d) Rarely
   e) Never

   Initial Code : Interest in others' opinion

9) Which machine translation tool do you mostly use ?
   a) Google Translation
   b) Microsoft Bing Translation
   c) Yandex Translation
   d) Baidu Translation
   e) Youdao Translation
   f) Other (please specify)

   Initial Code : Types of translation tools

10) Sentiment analysis tool can determine the attitude (positive or negative) of writers, which can be useful for analyzing a huge amount of comments. Are you willing to use this tool for analyzing your interest topic in social media, such as product comments or restaurant review, etc?
    a) Very likely
    b) Likely
    c) Neither likely nor likely
    d) Unlikely
    e) Very unlikely

Initial Code : Attitude towards sentiment analysis

REFERENCES

[1] 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.

[2] H. Peng, E. Cambria, and A. Hussain, "A review of sentiment analysis research in chinese language," *Cognitive Computation*, vol. 9, no. 4, pp. 423–435, 2017.

[3] "machine translation (1950s) linguistics." *Dictionary of Theories*, p. 323, 2002. [Online]. Available: http://ezproxy.uow.edu.au/login?url=https://search-ebscohost-com.ezproxy.uow.edu.au/login.aspx?direct=true&db=f6h&AN=40423164&site=eds-live

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

[5] 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.

[6] A. Sethi, "A review paper on levels, types & techniques in software testing," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 7, 2017.

[7] 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.

[8] Y. Cao, Z. Q. Zhou, and T. Y. Chen, "On the correlation between the effectiveness of metamorphic relations and dissimilarities of test case executions," in *Quality Software (QSIC), 2013 13th International Conference on*. IEEE, 2013, pp. 153–162.

[9] C. Henard, M. Papadakis, M. Harman, Y. Jia, and Y. Le Traon, "Comparing white-box and black-box test prioritization," in *Software Engineering (ICSE), 2016 IEEE/ACM 38th International Conference on*. IEEE, 2016, pp. 523–534.

[10] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*. Springer, 2012, pp. 415–463.

[11] A. Perrin, "Social media usage," *Pew research center*, pp. 52–68, 2015.