

# Metamorphic Testing in cross-language Sentiment Analysis of 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—**

**Index Terms—sentiment analysis, machine translation**

## I. INTRODUCTION

### REFERENCES

[Sen14] Mark Senn. *Using L<sup>A</sup>T<sub>E</sub>X for Your Thesis*, 2009 (accessed February 3, 2014). <http://engineering.purdue.edu/~mark/puthesis>.

This article was published in the New York Times.

Social Media is becoming more and more popular. There are lots of text comments on different discussion topics every day. It would be impossible to analyse 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 culture 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. The aim is for people to be able to understand different language speakers' attitudes (positive or negative), emotions and opinions.

[1]

Sentiment analysis is useful for analyzing a huge amount of data relating to personal opinions. It can be use in the e-business context. For example, a business manager can analyse customers' attitudes; whether they like or dislike their product or service. Governments can also use sentiment analysis to analyze citizens' perspectives.

In a word, sentiment analysis is coming into widespread used.

In this semester, I am firstly focus on conducting literature review about the current state of Chinese Sentiment Analysis research.

A review article published last year has been identified as relevant (Haiyun,2017). As Dr. Haiyun mentioned that English

language sentiment analysis research has undergone major developments in recent years.

Chinese sentiment analysis research has not evolved significantly.

Today, a lot of English language sentiment analysis theories have been developed.

Also, lots of translation tools are available.

However, nobody working in the two fields of research is undertaking non-English sentiment analysis.

So, the purpose of this research is to achieve a method for finding which machine translation 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 analyze their own language.

However, Chinese research does not sufficiently represent all of the non-English speaking world.

So, more information in Hindi, Arabic, Spanish, Portuguese and Japanese needs to identify. As a result, this research is proposed to create a testing model for finding the best-combined mode between English sentiment analysis tools and translation tools. There are couples of search terms used for this literature review, such as sentiment analysis, text mining,

machine translation, text segmentation, text classification, machine learning, deep learning, neural network, metamorphic testing and adaptive random testing.

In this literature review, it has 2 key journals, which are IEEE and ACM.

The key author is Zhiquan (George) Zhou.

All of reference article will come from books, research articles, review articles and website in recent 10 years, this research consists of three components.

Firstly, measuring machine translation service quality based on sentiment analysis results from both sides.

Secondly, testing sentiment analysis service quality.

Thirdly, finding the best compound mode for machine translation service and sentiment analysis service.

In the following statement, I will talk about Testing Method based on Metamorphic Testing, Machine Translation testing, Sentiment Analysis Testing and literature review of Experiment Support.

## II. MACHINE TRANSLATION

I have found two research articles about testing modeling of machine translation. Somers (2005, p. 127 -133) argue that using around trip translation method to detect the quality of machine translation. For example, testing English to Chinese translation tool. Firstly, using English to Chinese translation tool translate test data to Chinese. Secondly, translation Chinese translated data back to English. Finally, compare the similarity for two English data set. Another article is about using third-party language to test the quality of machine translation (Pesu, 2017). For example, if testing English to Chinese translation tool. Firstly, random choice an Intermediate third-party language. Secondly, translation English test data to the third-party language, after translation third-party language to Chinese. These two steps are only finished one path. Another path is a translation in English directly to Chinese. In the end, the comparison the two paths of results similarity. In this article, the main finding is Google Translate is the best machine translation. In addition, there is a tendency to be produced better results in European languages, which use ANOVA Statistics method getting this conclusion. In my experiment, I also got Google Translator is the best machine translation compare with Yandex and Baidu. The three different machine translation testing modeling, we can call the three different Metamorphic Relations. So, I have found a research article

Cao 2013

about the effectiveness of metamorphic relations. Also, when I finished reading all of the details, I have found Two-way ANOVA is also suitable for part 3 of my research (finding the best compound mode for machine translation service and sentiment analysis service).

I think that the disadvantages of those three-testing modeling will be involved some of the noise. It is clear the Round-Trip will be involved translation back path noise. Using the third language will be involved in third language translation path noise. My modeling will be involved sentiment analysis tools' noise. However, my modeling does not involve any noise from translation tool. I think my modeling is better than the other two testing modeling. I am total from another aspect to test translation tool.

## III. SENTIMENT ANALYSIS TESTING

In this semester, I have finished experiment for testing Google, Yandex and Baidu machine translation service, also finished experiment for comparison Google and Baidu sentiment analysis tools. For the sentiment analysis, I only finish testing the overview performance. Next semester, I have prepared to test sentiment analysis tools by different components. So, I have found a book (Liu, 2012) about sentiment analysis.

According to Liu's book that text sentiment analysis roughly have 5 steps. 1. Texting segmentation for the separate sentence to a list of keywords 2. Filtering for removing unnecessary keywords that will not add value to sentiment analysis, Such as is, but, it and so on. 3. Finding the basic words for Convert all infections to their root words. 4. Making Features for Using the root words as features to indicate the positiveness or negativeness 5. The last step is Classifier for the train a classifier to predict positivity. I will base on these five steps to testing sentiment analysis in next semester.

Testing Step 1 (Sentiment Analysis) Ramos (2003, p. 133 - 142) claims that Tf-idf, which is an abbreviation for the term Frequency Inverse Document Frequency, is one of words ranking calculation method and a kind of numerical statistic designed to reflect how important a word is to a collection or a corpus of documents. I think that Tf-idf is useful for testing sentiment analysis first step (texting segmentation). I can look for what is different between Baidu texting segmentation Ranking Group and Google texting segmentation Ranking Group in same test data. If these two rankings are different, I may identificate the quality for texting segmentation.

Testing Method Accounting to Sethi (2017) said that 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. In my research project, I will focus on Functional Testing.

Metamorphic Testing The proposed research is based on Metamorphic Testing. Metamorphic Testing is for testing function correctness. I have read a research article (Zhou, 2016), I think this article is well-explained for what is Metamorphic Testing. My testing modeling is based on Metamorphic Testing. So, this article is useful for me. As George (2016) said, Metamorphic testing (MT) is a property-based software testing method developed for automated test case generation and automated result verification, based on the effects of some expected properties of the target program. These properties, recognized as metamorphic relations (MRs), serve as essential relations among the inputs and outcomes of multiple executions of the target program. I make a example here to explain Metamorphic Testing. For instance, Test calculator function correctness will be shown if input  $1 + 1$ , and the result will be 2. In this example we can easily make the judgment, correct or not. However, we will not easy to make this judgment quickly if input  $\sin(3.7)$ , and the calculator gives a output. 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 my example, the relation is " $=$ ". I think the advantage of Metamorphic testing (MT) method, which can automate result verification and test case generation. The disadvantage is that cannot detect memory leak or some others insensitivity failure situation. However,

Metamorphic Testing is appropriate for testing translation tools and sentiment analysis tools. Effectiveness of Metamorphic Relations I have read an article which was written by Zhou in 2013. The main purpose of reading this article is to find the best one among the three testing modelings. 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. I think 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 calculation the program’s distance. In this research project, my testing modeling have not source code available. So, I have found another research article (Henard,2016)talking about the difference between black-box testing and white-box testing.

White-box VS Black-box 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

#### IV. EXPERIMENT SUPPORT

Stanisic (2015) have done a review article about experiment workflow. This review article is not related to my research topic, But very useful for improve research workflow, especially for writing experiment note. Git is version control system, which can rollback to older version of you writing. Emacs org mode is so powerful for you writing. Org-ref is citations, cross-references, indexes, glossaries and Bibtex utilities for org-mode. Org mode also can easier convert to latex format and html format. Even we can write your own emacs lisp for needed function. In my research, I need collect lots of data. So, I have read a website give me some ideas about how to manage experiment data. My personal opinion, if you write your own programs, such as using c++ or Python, I think the best way is using CSV format. Reasons: In C++ or Python, there are a lot of free libraries for reading and writing CSV format. In C++, I have found a library, which name is libxl(<http://www.libxl.com/>) for reading and writing Excel file but not free. In Python, there is a open source Excel library, which name is openpyxl, but I have found this library are not suitable for large amount of data operation, because very slow. Openpyxl is a Python library for reading and writing Excel 2010 xlsx/xlsm/xltx/xltn files. However, Microsoft excel have lots of build-in function for operation data. CSV can convert back to Excel format, if Microsoft

Excel build-in function is suitable for your research project, which is better way, it is not necessary for us to rewrite function by self. For example, you write your own coding for collect data in CSV file, after you convert CSV file to excel for draw graphic. When you find Excel have not suitable build-in function for you. You can convert back to CSV. To sum up, several research literature indicate sentiment analysis is coming into widespread used, but the evidence of the measuring machine translation service quality is critical point, it lies in best compound mode for machine translation service and sentiment analysis service. Functional Testing as main test methods, Metamorphic Testing in cross-language sentiment analysis service are needed. Metamorphic Testing is appropriate for testing translation tools and sentiment analysis tools. Several studies have indicated that black-box testing, and white-box testing may play a part in the effectiveness of metamorphic relations.

#### ACKNOWLEDGMENT

#### REFERENCES

- [Sen14] Mark Senn. *Using L<sup>A</sup>T<sub>E</sub>X for Your Thesis*, 2009 (accessed February 3, 2014). <http://engineering.purdue.edu/~mark/puthesis>.  
This article was published in the New York Times.