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*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

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

I. Introduction

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 some topic 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 [**?**]. 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, lots of Machine Translation tools are available, such as Google translation, Bing translation and Yandex translation. Manchine Translation uses computational linguistic programs and natural language processing theory [**?**].

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 document aimed 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.

II. Literature Review

This research consists of three components. Firstly, measuring machine translation service quality. Secondly, testing sentiment analysis service quality. Thirdly, finding the best compound mode for machine translation service and sentiment analysis service. As a result, this section will fouced on machine translation and sentiment analysis.

A. Machine Translation

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

a) Around Trip Translation Method: [2] As Somers argue that they establish an around trip translation method for 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.

b) Translation Method: 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. Method

IV. 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 %

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

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

Ranking	IQR	Mean	Median	Q1	Q3	lowerExtreme	upperExtreme
10	0.3426533	0.2404781	0.1789910	0.0489987	0.3916520	-0.4649812	0.9056319
20	0.3817485	0.2949725	0.2489580	0.0878135	0.4695620	-0.4848092	1.0421847
30	0.4504405	0.3980354	0.3808120	0.1516095	0.6020500	-0.5240512	1.2777107
40	0.5255830	0.5123376	0.5385910	0.2461685	0.7717515	-0.5422060	1.5601260
50	0.5368080	0.5712128	0.6188395	0.3101650	0.8469730	-0.4950470	1.6521850

Mean Slope	Median Slope
0.008788344	0.0116933

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