A Comparative Analysis of Translation Performance: ChatGPT vs. Google Translate

Abstract

This research project aims to assess and compare the translation capabilities of two popular language processing systems, ChatGPT-3 and Google Translate. The study employs the standardized BLEU algorithm to evaluate the quality of translations provided by each system. The objectives are to determine if ChatGPT outperforms Google Translate in terms of translation accuracy, identify potential variations in their translations, and examine the ease of translating different languages irrespective of the software used. Additionally, the study aims to establish whether there is a statistically significant difference between the two services. The hypothesis suggests that, due to ChatGPT's extensive dataset, it will yield higher BLEU scores compared to Google Translate. A large variety of phrases are used as test cases, each translated twice. The response variable measured is the BLEU score, and a two-way ANOVA is performed on the BLEU scores.

I. Background and Significance

Language translation has become an essential aspect of our increasingly globalized world, enabling effective communication and collaboration across linguistic barriers. With the advancement of technology, various machine translation systems have emerged to facilitate efficient and accurate translations. Two prominent translation methods widely used by individuals and organizations are ChatGPT and Google Translate.

Language translation plays a vital role in facilitating global communication. ChatGPT utilizes extensive data to generate contextually relevant translations, while Google Translate relies on algorithms and vast datasets. Despite their widespread use, there is limited comprehensive research comparing the performance of these two translation methods.

This research project aims to compare the translation performance of ChatGPT and Google Translate using the standardized BLEU algorithm. The objective is to evaluate the accuracy and effectiveness of these translation methods and identify any variations in their translations. The study seeks to fill the gap in comprehensive research comparing ChatGPT and Google Translate and provide valuable insights for individuals, businesses, and organizations in selecting the most suitable translation option. Additionally, understanding the strengths and weaknesses of these methods will contribute to improving cross-linguistic communication and advancing the field of machine translation.

The specific goals of this study are to assess whether ChatGPT outperforms Google Translate in terms of translation accuracy and to explore any potential variations in their translations. By comparing these widely used translation methods, we aim to provide guidance for selecting appropriate translation tools and contribute to the advancement of machine translation.

II. Statistical Methodology

The experiment was conducted using Google Translate and ChatGPT to translate a random list of phrases ranging in complexity from the original English phrase to the language and back again. The two phrases were then compared using Python code that compared the two languages with the BLEU algorithm. This algorithm compares the two different English phrases and spits out a score from zero to one, with a one being a perfect score and a zero being completely different. For the Google Translate phrases, we were able to use a total of 6870 phrases and for the ChatGPT we had to manually translate a limited number of 199 phrases into the AI. Please note that the ChatGPT translated phrases were translated on different computers to ensure ChatGPT isn't looking back into the old prompts you gave it.

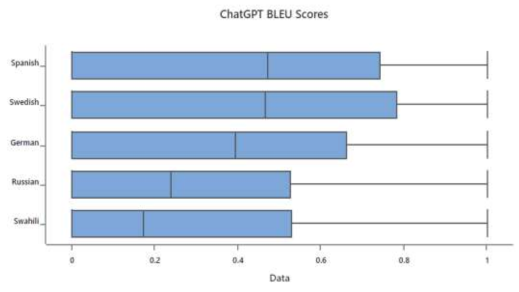
The response variables that we are measuring is the BLEU score. The input variable is the act of translation across the five languages that we selected. We decided to do a variety of different languages that are not just Latin based; we chose to use Spanish, Swedish, German, Russian, and Swahili for each software language model. Once we acquired all the BLEU scores, we performed the two-way ANOVA test on our data. Due to us having two factors: one being the software used to translate and the other being the language used, we need to analyze based on those parameters. This is used to illustrate the differences seen in the software and the languages. To demonstrate these differences, we needed a p-value of less than 0.05 to be considered in the comparisons. We used box and whisker plots to visually see if there is a difference between the two software.

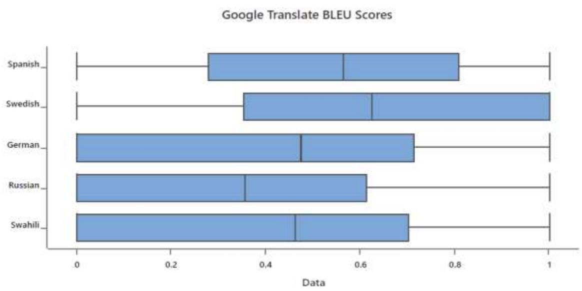
III. Results

Regarding the experiment, the two-way ANOVA analysis using the Tukey comparison method included the interaction between the software and languages as well as language and software on their own. The results indicate that there is not sufficient evidence (p-value = 0.062) to support a statistically significant interaction between software and language. However, there is sufficient evidence (p-value = 0.000) to conclude that there is a significant difference between the software and between the languages used. This means that we can reject the null hypothesis that there is no difference between the two software.

Based on the Tukey Method for 95% confidence comparison, the average BLEU score of Google Translate (0.471183) is statistically higher than that of ChatGPT (0.379080). These findings suggest that Google Translate is a more effective means of translating languages back and forth for the languages tested in this study.

Additionally, when considering the language factor irrespective of the software used, the Tukey comparison method revealed that Spanish (0.515841) and Swedish (0.490696) both yielded higher BLEU scores compared to German (0.426223), Swahili (0.362340), and Russian (0.330557).





IV. Discussion

The objective of this research project was to assess and compare the translation performance of ChatGPT and Google Translate using the BLEU algorithm and determine if ChatGPT outperforms Google Translate in terms of translation accuracy. Additionally, the study aimed to explore any variations in translations and establish whether there is a statistically significant difference between the two services as well as to see if certain languages were easier to translate regardless of the service.

The results of the two-way ANOVA analysis indicated that there was not sufficient evidence to conclude a significant interaction between software and language (p = 0.062). However, a significant difference was observed between the software (p = 0.000), with Google Translate yielding higher BLEU scores on average than ChatGPT. These findings suggest that, for the languages tested in this study, Google Translate is a more effective translation tool.

Although efforts were made to ensure the normality of the data, the analysis revealed that the distribution of the residuals did not meet the assumption of normality. This violation suggests that the data may not follow a perfectly normal distribution. However, it is important to note that the ANOVA analysis is robust against violations of the normality assumption, and the Type 1 error rate remains close to the specified alpha level. Therefore, the findings of the ANOVA analysis can still be considered reliable for drawing conclusions and making inferences.

Nonetheless, future studies may consider exploring alternative non-parametric tests or robust methods that are specifically designed to handle non-normal data. This would provide a more comprehensive understanding of the translation performance of ChatGPT and Google Translate.

In the larger scope of previous research on language translation, this experiment contributes by providing insights into the performance of two popular translation methods. By utilizing the BLEU algorithm, the study addresses the need for comprehensive research comparing the translation capabilities of ChatGPT and Google Translate. The findings help inform individuals, businesses, and organizations in selecting the most suitable translation method for their needs.

It is important to acknowledge the limitations of this study. Firstly, the analysis was limited to a specific set of languages and may not be representative of all language pairs. Additionally, the study focused on the BLEU algorithm as the sole evaluation measure, and future research could explore the use of additional metrics to provide a more comprehensive assessment of translation quality. Furthermore, expanding the sample size and including a wider range of phrases and language pairs would enhance the generalizability of the findings. Additionally, if we had access to ChatGPT-4 the language translation might be better than ChatGPT-3.

In conclusion, this comparative analysis demonstrates that Google Translate outperforms ChatGPT in terms of translation accuracy based on the BLEU scores for the languages tested. Moving forward, further research can build upon these findings to explore other language pairs, utilize additional evaluation metrics, and enhance our understanding of machine translation.

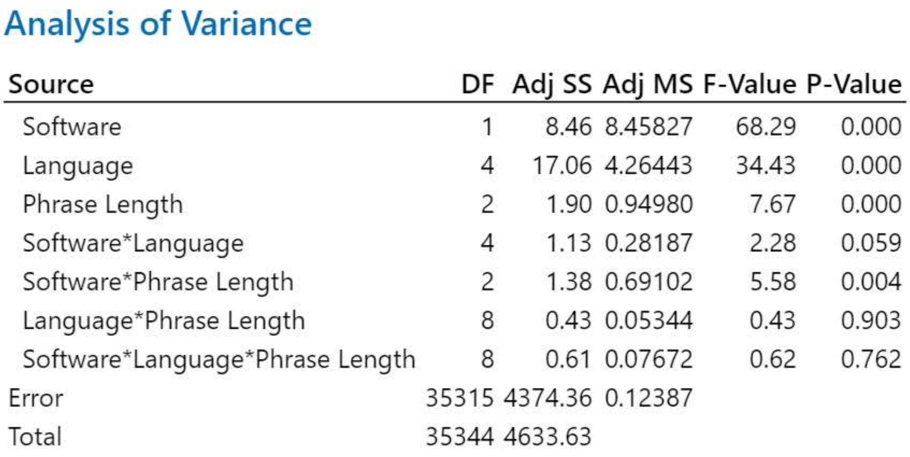
V. References

Parlett-Pelleriti, Aaron R. Caldwell, Daniël Lakens, and Chelsea M. Chapter 11 Violations of Assumptions | Power Analysis with Superpower. Aaroncaldwell.us, aaroncaldwell.us/SuperpowerBook/violations-of-assumptions.html.

“Nltk - Calculate BLEU Score in Python.” Stack Overflow, stackoverflow.com/questions/32395880/calculate-bleu-score-in-python.

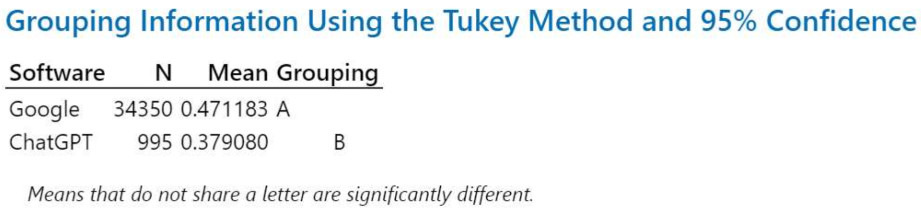
World, Woman of the. “What Is BLEU Score? (and How Does It Affect Translation?).” Asian Absolute UK, 28 Jan. 2021, asianabsolute.co.uk/blog/2021/01/28/what-is-bleu-score/.

Appendix



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Software	1	8.46	8.45827	68.29	0.000
Language	4	17.06	4.26443	34.43	0.000
Phrase Length	2	1.90	0.94980	7.67	0.000
Software*Language	4	1.13	0.28187	2.28	0.059
Software*Phrase Length	2	1.38	0.69102	5.58	0.004
Language*Phrase Length	8	0.43	0.05344	0.43	0.903
Software*Language*Phrase Length	8	0.61	0.07672	0.62	0.762
Error	35315	4374.36	0.12387		
Total	35344	4633.63			



Grouping Information Using the Tukey Method and 95% Confidence

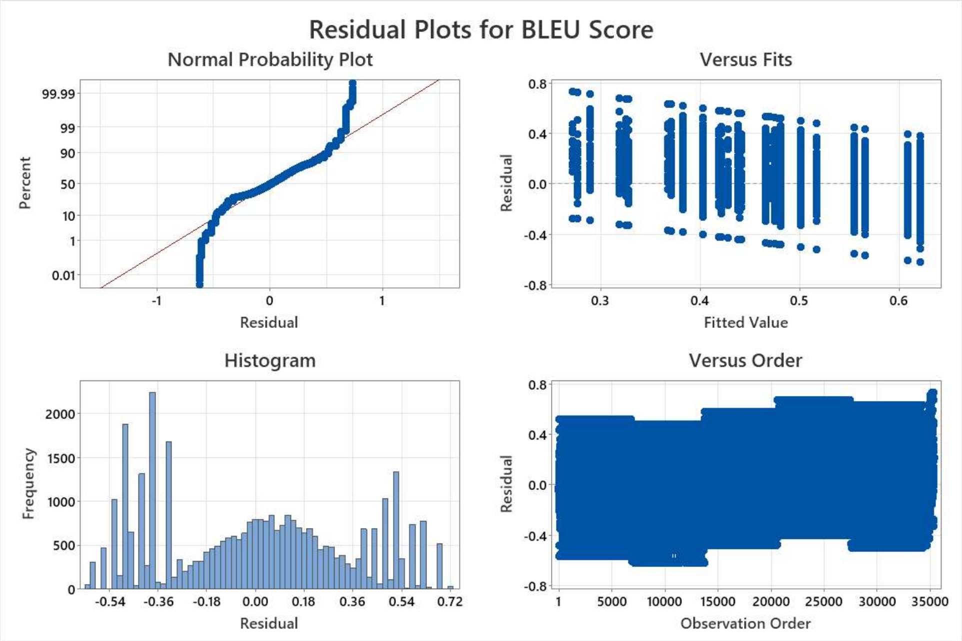
Software	N	Mean	Grouping	
Google	34350	0.471183	A	
ChatGPT	995	0.379080		B

Means that do not share a letter are significantly different.

Grouping Information Using the Tukey Method and 95% Confidence

Language	N	Mean	Grouping		
Swedish	7069	0.515841	A		
Spanish	7069	0.490696	A		
German	7069	0.426223		B	
Swahili	7069	0.362340			C
Russian	7069	0.330557			C

Means that do not share a letter are significantly different.


Residual Plots for BLEU Score
Normal Probability Plot
Percent
Residual
Versus Fits
Residual
Fitted Value
Histogram
Frequency
Residual
Versus Order
Residual
Observation Order