Experiments and Perceptions in Machine Translation

Hsiao-Chien Wei, Li-Ci Chuang, Mu-Hua Hsu, Su-Tien Lee, Xue Han
Matthew A. Lanham

wei365@purdue.edu; chuangl@purdue.edu; hsu269@purdue.edu;
lee3836@purdue.edu; xue161@purdue.edu

spf.io | PURDUE UNIVERSITY | Krannert School of Management

ABSTRACT

We examined off-the-shelf machine translation (MT) models provided by Google and Microsoft platforms to gauge how well they translate. Platforms such as Google and Microsoft offer a way to build customized models. However, the translation quality of these platforms varies based on customer survey research. We designed and iterated through several experiments to find solutions to improve translations by preparing the trained datasets in a certain fashion and achieve higher translation accuracy at the same time.

INTRODUCTION

In the post-pandemic era, we have seen global conferences go from face-to-face to online. However, it is difficult to hold events that require Zoom translation services for multiple broadcasts. If we can reach a higher accuracy and a broader application of machine translation, we can better meet the market's need for real-time translation in meetings, providing every user a decent experience when attending these events and enabling human to cross borders.

Source: Mordor Intelligence

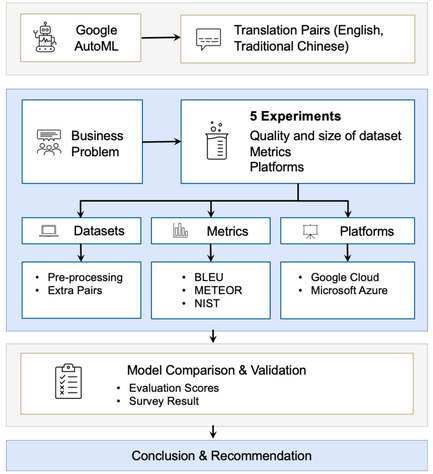
Fig 1. Machine Translation Market Development Prediction

We used the AutoML Adaptive Machine Translation to train our model with datasets, which consisted of pairs of Biblical texts in English and Traditional Chinese, respectively. We tried different numbers of pairs on different platforms to see how the performance improved based on machine translation metrics and human evaluation.

RESEARCH OBJECTIVES

- Which platform performs better with customized datasets?
- What's the relationship between size and diversity of trained datasets and the results of translation?
- Which evaluation metrics of machine translation is more consistent with human understanding?

METHODOLOGY



STATISTICAL RESULTS

We conducted five experiments in total with varied datasets within Google platform as well as the same datasets across Google and Microsoft Azure. By comparing evaluation scores and the survey rankings, we found out the inconsistency among the platform scores, self-calculated scores, and human evaluations.

Experiment	Base BLEU	Customized BLEU	Improvement
MS Extra	35.66	38.83	8.89%
Google Raw	45.92	47.96	4.44%
Google 1000	45.80	47.81	4.39%
Google Extra	45.81	47.03	2.66%
Google Test	45.55	46.49	2.06%

Fig 2. Model BLEU Results

Trial	Dataset
Google Test	Raw Dataset1 (Dataset1)
Google 1000	Mixed (500 Dataset1 + 500 Dataset2)
Google Raw	Dataset1+Dataset2 (Duplicates Dropped)
Google Extra	Dataset1+Dataset2 (Duplicates Dropped) + Extra Pairs
MS Extra	Dataset1+Dataset2 (Duplicates Dropped) + Extra Pairs

Fig 3. Trial Data Description

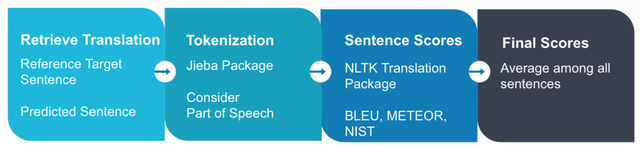


Fig 4. Self-Calculated Scores

We used Friedman Test to further test whether there are different perceptions among translations within the Google platform and across platforms for both Christians and non-Christians. The critical value with 95% confidence level for our sample is 7.81, so we would view the result as statistically significant if its corresponding Chi-square value is greater than 7.81.

	Within Google Platform	Across Platform
Christian	28.82	26.37
Non-Christian	1.9	11.39

Fig 5. Friedman Test results

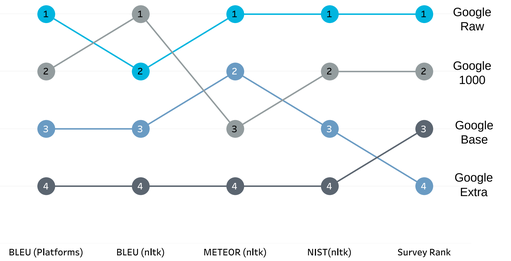


Fig 6. Model Performance within Google with Different Metrics

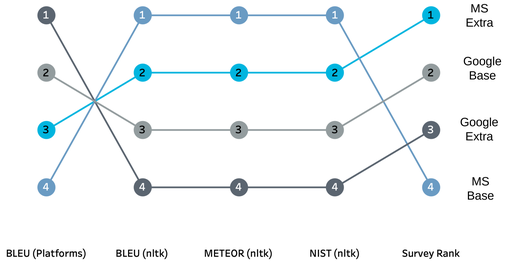


Fig 7. Model Performance across Platform with Different Metrics

EXPECTED IMPACT

A well-trained customized translation model can help our client's customers contribute cost savings by decreasing the needs to hire human translators for traditional post-editing services. If our client fully turns to Spf.io with one year's subscription, there will be about 90% economic cost savings and about 98% in time savings.

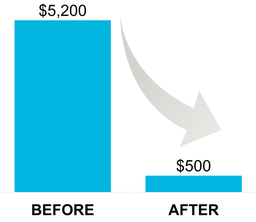


Fig 8. Impact on Economic Cost

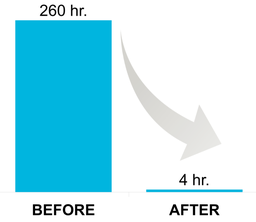


Fig 9. Impact on Time Cost

CONCLUSIONS

Platforms

For small to medium-sized companies, we recommend staying with Google AutoML. By doing so, they would not need to collect large datasets and there will be no platform switching costs. On the other hand, Microsoft can be utilized to serve large-sized customers, not only for its better performance in cross-platform experiments but also for its stability.

Data Preparation

Larger data brings about better translation performance. However, the size of dataset is not as important as the consistency with data source. Furthermore, in the same source of data, we encourage diversity in topics. Thus, for small to medium-sized companies with a limited budget, there should be a tradeoff between large size and high-quality datasets. In such cases, we would recommend pursuing quality.

Evaluation Metrics

Either one could be used to evaluate translation performance. However, we should not rely on the platform's calculated scores only, and human evaluation should be taken into consideration as well.

ACKNOWLEDGEMENTS

We would like to thank Professor Matthew Lanham and our industry partner Spf.io for this opportunity, their guidance, and support on this project. We would also like to thank the survey respondents who took valuable time away from their day jobs to participate in this work.

informs