Intelligently Ordering Machine Translation Seed Data to Improve Local Language Translation

Amaan Ansari, Devansh Batra, Jai Woo Lee, Paul Chen, Gagan Pahuja, Manideep Sharma, Daniel Whitenack, Matthew A. Lanham
Purdue University, Krannert School of Management
ansari4@purdue.edu; batra17@purdue.edu; lee3999@purdue.edu; chen3876@purdue.edu; gpahuja@purdue.edu; sharm536@purdue.edu; lanhamm@purdue.edu

ABSTRACT

There is a critical need for optimizing the seed data needed for creating a widely acceptable machine translation model for low-level local languages. This research addresses the seed data concern by determining an optimized order of seed data which results in both more accurate and quicker translations as compared to a random order. This is achieved by dividing data from large translation project into various combination of test and train sets and achieve a BLEU score on the test data in the least amount of time and with the least number of iterations.

INTRODUCTION

- Translation cost is a significant limiting factor on the pace and availability of translated important content.
- There are only handful of available methods to achieve the accuracy with minimum seed data usage.
- This project will demonstrate the *'high-accuracy low-data dependent'* algorithm that can be generalized and scaled across different languages to create effective translation to low-level local languages.

RESEARCH OBJECTIVES

Our research focuses on answering the following questions:

- What are the most important factors that play a role in optimizing the seed data for language translation?
- What kind of role does supplemental data pertaining to similar semantic domain play in optimization?
- How can business across world utilize this research to reduce translation cost?

METHODOLOGY

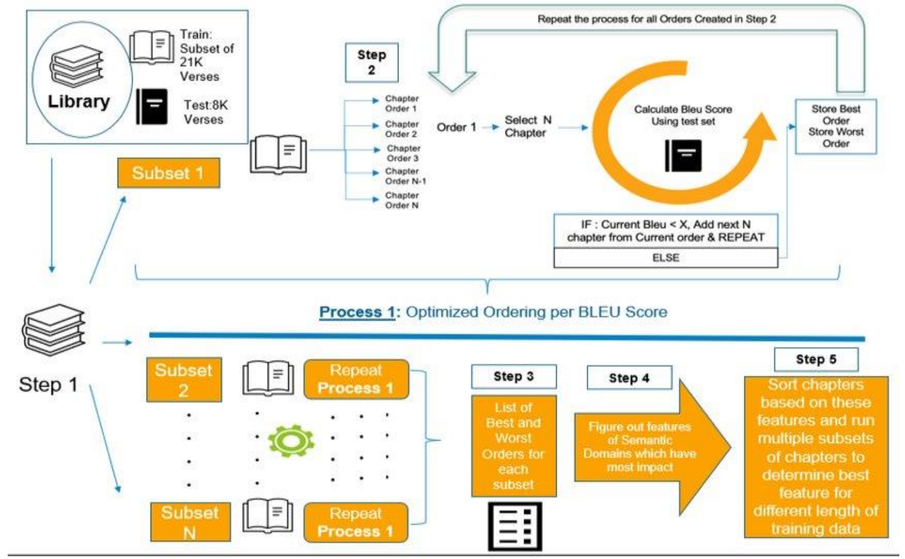
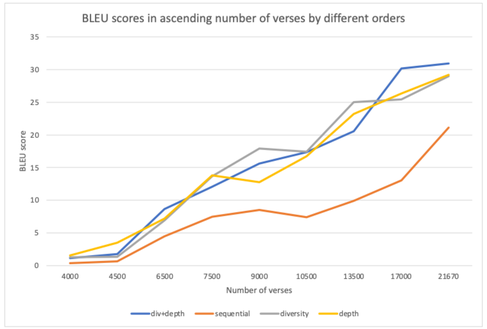


Fig 1. Optimization Design

STATISTICAL RESULTS



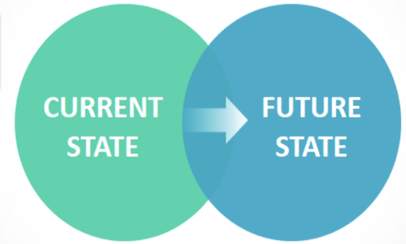
- As we increased the number of verses provided into the model, the BLEU scores from three different orders outperformed sequential order significantly (using t-test)
- Three models based on the suggested orders reached a similar BLEU score to that of sequential 21,670 verses, using 38% (10,500 verses)- 52% (13,500 verses) less number of verses

BUSINESS IMPACT

SIL is **currently engaged in 1600+ language projects** around the world

Have **direct costs to SIL of $1.3M** per project

Lasts up to 15 years



SIL will **expand the translation** projects to up to 2000 by 2022

Our MT models in loop with human translators will lead to **potential savings** of **$600k per project** for SIL Or **$6M for 10-15 large projects**

Reduce time to completion by 7-8 years.

A set algorithm for translations engines, which will allow client to optimize the translation seed data

CONCLUSIONS

- For low resource languages, i.e., less training data, depth of semantic domain is more important to achieve a higher accuracy(better translation).
- Once training data is sufficient, combination of depth and diversity of semantic domains plays a significant role in boosting the accuracy (BLEU Score).
- Statistical Result of our MT model (Javanese translation) shows that if the training data is kept as low as 5000 text sentences, then depth of the semantic domains boost the BLEU score to ~7, however with larger training set, combination of depth and diversity of semantic domain can boost the score to ~30 without any supplemental data.
- We believe that adding the external supplemental data will boost the accuracy rate even more. However, that data should have sufficiently high or comparable diversity and depth score with Bible text.

ACKNOWLEDGEMENTS

We thank our industry partner Daniel Whitenack for his trust, support and encouragement while approaching this problem. We also thank Professor Matthew Lanham for constant guidance on this project.