Intelligently Ordering Machine Translation Seed Data to Improve Local Language Translation

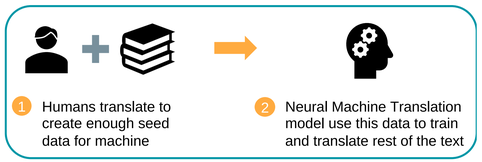
Amaan Ansari, Devansh Batra, Jai Woo Lee, Paul Chen, Gagan Pahuja, Manideep Sharma, Daniel Whitenack, Matthew A. Lanham
Purdue University, Krannert School of Management
ansari4@purdue.edu; batra17@purdue.edu; lee3999@purdue.edu; chen3876@purdue.edu; gpahuja@purdue.edu; sharm536@purdue.edu; lanhamm@purdue.edu

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

Status-quo Translation of Low-resource Language



1. Humans translate to create enough seed data for machine
2. Neural Machine Translation model use this data to train and translate rest of the text

Translation Cost	Human Error	Amount of Train Data Needed
HIGH	HIGH	LARGE

- 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 for 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?

LITERATURE REVIEW

While all relevant studies focus on only machine translation or optimization (e.g. dynamic selection of seed data), none of them focused on optimizing seed data through ordering them differently based on semantic aspects.

Study	Ordering based on semantic meaning	Machine Translation	Optimization
Our methodology (2022)	YES	YES	YES
Liwei Wu, Shanbo Cheng, Mingxuan Wang, Lei Li (2021)		YES	
Xinyi Wang and Graham Neubig (2019)		YES	YES
Laura Martinus and Jade Abbott (2019)		YES	YES
Marlies van der Wees, Arianna Bisazza, and Christof Monz (2017)		YES	
Ranathunga, S., Lee, E. S. A., Skenduli, M. P., Shekhar, R., Alam, M., & Kaur, R. (2021)		YES	YES

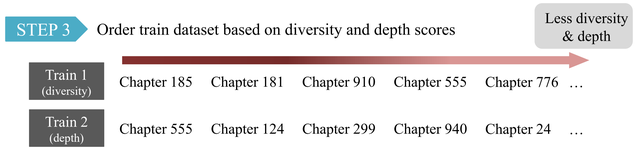
METHODOLOGY

STEP 1 Create and clean parallel corpora

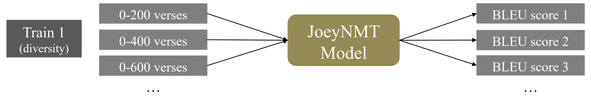
	Book	Chapter	Verse	Sequence	English	Javanese
0	GEN	1	1	1	In the beginning God created the heavens and t...	Ing jaman kawitan Gusti Allah nitahake langit ...
1	GEN	1	2	1	Now the earth was formless and empty, darkness...	Anadene bumi mau campur-bawur lan suwung, sega...
...	...	...	...	...	...	...
29952	REV	22	20	1144	He who testifies to these things says, "Yes, I...	Kang nglairake paseksen bab iki mau kabeh ngan...
29953	REV	22	21	1144	The grace of the Lord Jesus be with God's peop...	Sih-rahmate Gusti kita Yesus Kristus nunggila ...

STEP 2 Get diversity & depth score for each chapter and split into train & test dataset

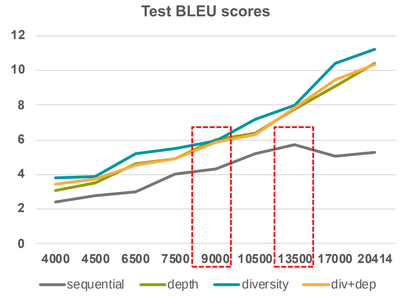
STEP 3 Order train dataset based on diversity and depth scores

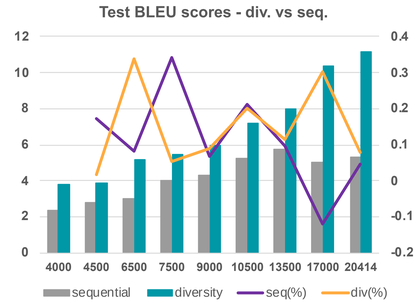


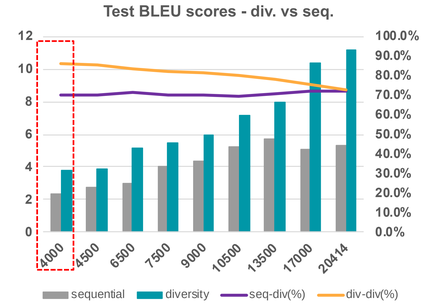
STEP 4 Run each subset of train set into JoeyNMT model and get test BLEU scores



STATISTICAL RESULTS







1. BLEU scores at 20,414

Diversity	11.2
Depth	10.44
Div+Dep	10.36
Sequential	5.3

2. Sent. used to get 6 BLEU

Diversity	8,000
Depth	8,000
Div+Dep	8,000
Sequential	13,500

3. Increase rate in BLEU

Diversity	12.7%
Sequential	9.3%

4. Percentage difference

Diversity	1.8% ~ 33.8%
Sequential	-11.8% ~ 34.1%

5. Increase rate in avg. div.

Diversity	-1.91%
Sequential	0.32%

6. Effect of avg. div. on BLEU

	Avg. div.	BLEU
Diversity	86%	3.81
Sequential	70%	2.38

16%p+ Avg. div. in 4,000 sent. → **60%+** BLEU

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

CURRENT STATE → FUTURE STATE

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 the client to optimize the translation seed data

- Based on our methodology and NMT model, we expect that our client can save as much as ~46% of human capital cost per project ($1.3M -> $0.6M)
- We also expect that the time to complete 10-15 large language projects can be reduced by as much as ~53% (15 years -> 7-8 years)

CONCLUSION

- Statistical result of our MT model shows that diversity of the semantic domain is the most significant parameter to achieving higher BLEU score for smaller as well as a larger training dataset
- Based on the diversity parameter, our model used about 40% fewer seed data (8,000 sent.) to accomplish the same BLEU score to the sequential parameter.
- We've found that using the most diverse sentences as early as possible in the early stage boosts not only the initial BLEU scores but also an increasing rate of BLEU scores in subsequent sets of sentences
  - In the first subset of 4,000 sent, having more diverse sentences by 16%p+ improved the initial BLEU score by 60% and the overall increasing rate by 3.4%p+
- 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 scores 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.