Intelligently Ordering Machine Translation Seed Data to Improve Low Resource Local Language Translation

Amaan Ansari, Devansh Batra, Jai Woo Lee, Paul Chen,

Gagan Pahuja, Manideep Sharma, Matthew A. Lanham

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

ansari4@purdue.edu; batra17@purdue.edu; lee3999@purdue.edu; chen3876@purdue.edu;

gpahuja@purdue.edu; sharm536@purdue.edu; lanhamm@purdue.edu

ABSTRACT

Neural machine translation (NMT) has become the state-of-the-art methodology for any Machine Translation task. However, there remain areas for improvement in the optimization of algorithms, hyperparameters, and the seed data itself for a more effective MT. Only a negligible fraction of the 7000+ currently spoken languages have sufficient text corpora to train MT models. This data scarcity results in systematic inequalities in the performance of MT across the world's languages. This research addresses the seed data concern to determine an optimized order of seed data which results in both more accurate and quicker translations as compared to a random/sequential order. This is achieved by dividing chapters from the English Bible into train and test data, then feeding different possible ordering combinations of train data one by one to identify which training order achieves a pre-defined BLEU score on test data in the least amount of time and with the least number of iterations (epochs). Once an optimized order is determined, a comparison is made between this order and the baseline order. Our findings suggest that the diversity, depth and their

combination are key in achieving the most accurate MT and provide a pathway to aaccelerate machine translation tasks.

1. INTRODUCTION

Language translation is one of the areas which can help communities across the world to exchange thoughts, ideas, and emotions and get a better understanding of nuances about the language and the culture itself. Many languages, especially in the secluded regions of the globe, are hard to translate due to lack of adequate texts, low resources, and unfamiliarity of the language with the general public. This project helps solve this problem by creating an algorithm through optimization of seed data that will determine the order and content of text translation and help a machine learn any language faster and create more accurate results with minimum amount of data.

Through this, we essentially help reduce the amount of effort required to translate large documents to provide the best input data for Neural Machine Translation (NMT) models. This research focuses heavily on understanding which factors play an essential role in efficient language translation using machine learning models. By understanding these, we can single out the logic for efficient translations and extrapolate them onto other languages for scalability and ease of communication. This research will be beneficial for the language and dialects that are not widely spoken and have few to no translations available. Hence, it will help such communities reach out and blend in with their high-resource language-speaking counterparts.

Using an inside out approach, a simple model is first built on a specific text corpus and then approaches such as Data Augmentation (Wang et al., 2019), Zero-Shot NMT (Liao ET AL., 2021), Target Conditioned sampling (Wang, Neubig, 2019) are implemented to efficiently train our model along with different combinations of the order and domains of seed data.

This project translates text from the Bible from English to Javanese. The goal is to find the effect of semantic domain order (chapter order) on the efficiency of language translation. We use the BLEU score as the standardized measurement of language translation accuracy across the texts. This gives us the chance to understand the effectiveness of the optimized order against the non-optimized ones and calculate the effect of optimization on the translation. It also helps us determine the perfect balance between computational power expenditure versus accuracy improvement. This research aims to provide a repeatable algorithm / methodology that can be used to translate low resource language at a higher level using minimum translated data by humans.

2. LITERATURE REVIEW

Most of the related papers demonstrated new methodologies for data selection and data adaptation to improve models' performance.

Martinus and Abbott (2019) [1] showed an approach to training and performing neural machine translation, particularly on low-resourced African languages from English. The experiment showed that the ConvS2S model with BPE tokenization performed better than the ConvS2S model with White-spaced tokenization by up to 9 BLEU. In comparison, the Transformer model with Word Piece tokenization performed even better up to 12 BLEU. The experiment showed that the performance is related to both the number of parallel sentences and the morphological typology of the language. While this paper provides a base guideline for model buildings and evaluation processes in neural machine translation on low-resource languages, using training data selectively rather than using them exhaustively to train models is not done and left to further analysis.

Wees, Bisazza, and Monz (2017) [2] performed dynamic data selection for NMT, which entails varying the selected subset of training data between training epochs. The experiment demonstrated

that gradually reducing the training size of the data improves the score consistently over conventional static data selection up to 2.6 BLEU. However, the research was done from English to German, a high-resource language. We need to test if this method still generates good results with low-resourced languages.

(Lee, et. al (2021)) [3] presents a detailed survey of research advancements in low- resource language NMT (LRL-NMT), along with a quantitative analysis aimed at identifying the most popular solutions. It provides a set of guidelines to select the possible NMT technique for a given LRL data setting. The research recommends 1) creating LRL resources (datasets and tools), 2) making computational resources and trained models publicly available, and 3) involving research communities at a regional level.

Wang, Neubig (2019) [4] constructed a sampling distribution over all multilingual data to minimize the low-resource language's training loss. Based on this formulation, they proposed an efficient algorithm, Target Conditioned sampling (TCS), which first samples a target sentence, and then conditionally samples its source sentence. They also conducted experiments that showed that TCS brings significant gains of up to 2 BLEU on three of four tested languages with minimal training overhead.

Wu, et. al (2021) [5] demonstrated that Language Tags (LT) are not only indicators for translation directions but also crucial to zero-shot translation qualities. A proper LT strategy could enhance the consistency of semantic representations and alleviate the off- target issue in z- shot directions. They conducted experiments which, by ignoring the source language tag (SLT) and adding the target language tag (TLT) to the encoder, the zero-shot translations could achievea +8 BLEU score difference over other LT strategies.

3. DATA

In collaboration with a non-profit organization (SIL International) which develops and document lesser-known languages, we got various Machine translation datasets listed below:

English Bible (NIV version): It contains 66 books and 1,189 chapters in total. Sample text in verses is shown below. We processed this raw data into the verses corresponding to each chapter which looks like the data in Table 1. This data is used for training the machine translation model.

English Bible Verses
on the twenty-first day of the seventh month, the word of the Lord came through the prophet Haggai:
Speak to Zerubbabel son of Shealtiel, governor of Judah, to Joshua son of Jozadak, the high priest, and to the remnant of the people. Ask them,
Who of you is left who saw this house in its former glory? How does it look to you now? Does it not seem to you like nothing?
But now be strong, Zerubbabel, declares the Lord. Be strong, Joshua son of Jozadak, the high priest. Be strong, all you people of the land, declares the Lord, ,and work. For I am with you, declares the Lord Almighty.
This is what I covenanted with you when you came out of Egypt. And my Spirit remains among you. Do not fear.
This is what the Lord Almighty says: little while I will once more shake the heavens and the earth, the sea and the dry land.
I will shake all nations, and what is desired by all nations will come, and I will fill this house with glory, says the Lord Almighty.

Table 1: Raw English Bible Text

Book	Chapter	Verse	Verses in English
HAG	2	1	on the twenty-first day of the seventh month, the word of the Lord came through the prophet Haggai:
HAG	2	2	Speak to Zerubbabel son of Shealtiel, governor of Judah, to Joshua son of Jozadak, the high priest, and to the remnant of the people. Ask them,
HAG	2	3	Who of you is left who saw this house in its former glory? How does it look to you now? Does it not seem to you like nothing?

Table 2: Processed English Bible Text for Training Purpose

Javanese Bible text: Like English Bible (NIV version), raw Javanese text is also pre-processed and aligned with its English counterpart to have line by line translated text for modeling purpose.

Book	Chapter	Verse	Verses in English
HAG	2	1	Ing dina kapisan ing wulan kaping pitu, tembung saka Gusti teka liwat Nabi Haggai:
HAG	2	2	"Ngomong karo Zerubabel, putrane Shealtiel, Gubernani Yehuda, marang Yosua bin Yozadak, Imam Agung, lan sisa-sisa wong. Takonipun,
HAG	2	3	'Sapa sampeyan isih kiwa sing ndeleng omah iki ing kamulyan? Kepiye carane sampeyan saiki? Apa ora kaya sampeyan ora seneng?

Table 3: Processed Javanese Bible Text for Training Purpose

Semantic domain mapping data: SIL has a repository of semantic domain mapping the one shown in table 4. This table was used to identify the semantic domains in English Bible data and further processed for depth and diversity for machine translation model input parameter.

word	category
moon	Moon
lunar	Moon
satellite	Moon
rise	Moon
set	Moon
sink	Moon

Table 4: Semantic Domain Mapping

4. METHODOLOGY

The goal of the semantic domain assignment (SDA) is to identify semantic domain distribution and quantify the degree of diversity and depth at chapter level. The English and Javanese sentences were aligned parallelly at chapter level and blanks were removed. Next, each word in the English

bible and in the semantic domain were vectorized using fast text library. The semantic domain words consist of 1783 domains such as moon, star, and air etc. Afterwards, cosine similarity was computed between each chapter and each semantic domain word by making use of the numba library. Based on the Euclidean distance, each word that was above a certain threshold was categorized into a corresponding semantic domain. Using this information, we created a semantic domain distribution table. Using this table, the degree of diversity and depth was computed for each chapter that can be used to put the chapter in an order for translation.

	1	2	3	4	5	6	7	8	9	10	...	1135	1136	1137	1138	1139	1140	1141	1142	1143	1144
Moon	7	2	0	0	0	0	0	1	1	0	...	0	0	0	3	0	0	1	0	2	1
Star	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Planet	7	6	0	1	0	12	14	9	7	3	...	5	6	0	1	5	6	2	2	3	0
Sun	0	0	0	0	0	0	0	0	0	0	...	0	0	0	2	0	0	1	0	2	1
Blow air	14	5	3	1	0	1	9	13	2	0	...	5	3	2	7	7	4	3	4	2	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Names of rivers	19	11	0	3	0	4	8	10	6	2	...	20	6	5	13	9	8	10	9	10	6
Name of a place	53	33	18	18	38	28	38	29	21	9	...	42	25	14	29	23	31	22	24	38	30
Names of animals	12	11	5	4	0	6	13	7	7	2	...	17	4	3	6	6	7	6	4	6	2
Names of buildings	3	6	0	5	0	4	2	1	3	3	...	17	5	3	8	7	11	6	5	17	6
Name of a thing	49	43	26	20	28	34	41	27	29	8	...	49	28	14	31	25	36	29	24	39	41

1783 rows × 1144 columns

Table 5: Semantic Domain Distribution Table

	chapter	diversity	english	javanese
184	185	1.00000	Listen, you heavens, and I will speak; hear, y...	“Kupingmu tilingna, he langit, aku arep catura...
180	181	1.00000	If you fully obey the Lord your God and carefu...	“Manawa kanthi temen-temen anggonmu ngestokake...
909	910	0.93333	When Jesus had finished saying all these thing...	Sawise Gusti Yesus mungkasi piwulange mau, ban...
554	555	0.93333	Blessed are those whose ways are blameless, wh...	Rahayu wong kang padha utama lakune, kang ngam...
775	776	0.93333	The word of the Lord came to me: “Son of man, ...	Nuli ana pangandikane Pangeran Yehuwah marang ...
...	...	...	...	...
528	529	0.30000	The Lord reigns, he is robed in majesty; the L...	Pangeran Yehuwah iku jumeneng raja, abusana ka...
555	556	0.28333	I call on the Lord in my distress, and he answ...	Kidung jiyarah. Sajroning karubedan aku sesamb...
552	553	0.21667	Praise the Lord, all you nations; extol him, a...	He sakehing bangsa, padha ngluhurna Pangeran Y...
585	586	0.20000	Praise the Lord. Praise God in his sanctuary; ...	Haleluya! Padha saosa puji marang Gusti Allah ...
569	570	0.15000	Praise the Lord, all you servants of the Lord ...	Kidung jiyarah. Ayo padha saosa puji marang Sa...

800 rows × 4 columns

Table 6: Parallel Data Ordered based on Diversity

After mapping the semantic domain, we created a methodology to find the semantic domain characteristics that determine the order of training text that would provide correct translation using least amount of data for low resource languages. To achieve this, we created four different orders based on the semantic domain parameters we wanted to test our results on i.e., diversity of domain and depth of domain.

To start with, we randomly split the data into 70% training and 30% testing. Training data was then arranged in four orders based on the semantic domain parameter after scaling semantic domain diversity and depth for each chapter between 0 and 1. In the first order, the training data was sorted in descending order of diversity scores, the second ordering was based on depth of semantic domain in each chapter of the training data, the third order was created by a combination of depth and diversity using multiplicative relation between two scores (diversity score * depth score). The fourth and final order were the natural orders that a human translator would use to translate any book i.e., from start to end. The next step was to split the training data into smaller

chunks to understand the effect of adding training data on the bleu score i.e., to determine the rate of increase in bleu score with addition of the training data from each metric (depth, diversity, depth*diversity). We split the training data into 4000, 4500, 6500, 7500, 9000, 10500, 13000, 17000 and 21670 verses. We keep on adding this data to see the rate of increase in bleu score for each of our metrics. To determine the bleu score, we used Joey NMT which has been described in detail in the 'model' section of this paper.

The BLEU scores for all our metrics are then compared against each other to infer the results. We look at each phase of our training data and see which metric gives a higher BLEU score in each phase.

5. MODEL

The preliminary model that we employed in this research is a Python-based neural machine translation model given by Joey NMT. This toolkit is made based on Pytorch. We chose this toolkit since Joey NMT supports both classic architectures such as RNN and Transformer and lets us to modify the parameters in comparison to Google's autoML. The advantages of neural machine translation is that it can handle large datasets with little supervision and has higher accuracy than traditional phrase-based machine translation. The neural machine translation uses neural networks to translate source language sentences to target language sentences with two main sections: an encoder neural network and a decoder neural network. The ability of the model to encode the source text into an internal fixed-length representation termed the context vector is critical to the encoder-decoder architecture. However, this design has a drawback in that it cannot be used to train large text sequences. To resolve this issue, we converted the source and target sentences to a line- by-line alignment format.

The critical configuration components that we developed for this model are as follows:

Scheduling	noam
Learning Rate	0.0003
Shuffle	False
keep_best_ckpts	3
eval_metric	BLEU
Learning_rate_factor	0.5
Learning_rate_warmup	1000

Table 7: Important Parameters

We set shuffle to false because the training dataset's order is critical to our research. Additionally, we utilize the BLEU score as an evaluation metric because it is the industry standard in the field of machine translation. For this study, we compared plateau and Noam and discovered that Noam has a slightly higher score.

6. RESULTS

The results show a clear advantage of ordering chapters based on diversity, depth, or combination compared to using chapters sequentially. As we increased the number of verses provided into the model, the BLEU scores from three different orders outperformed sequential order. While the sequential model reached the BLEU score of 21 using 21,670 verses, three models based on the suggested orders reached a similar BLEU score using 38% (10,500 verses) ~ 52% (13,500 verses) less number of verses. Exhausting all verses, the three orders helped the model reach the BLEU scores of almost 30. In addition, the BLEU scores between each order of the three and the sequential order are significantly different based on paired t-tests. Although we notice different slopes in many different ranges of verses among the three orders, the overall trends are almost the same, and BLEU scores are not significantly different according to the paired t-tests.

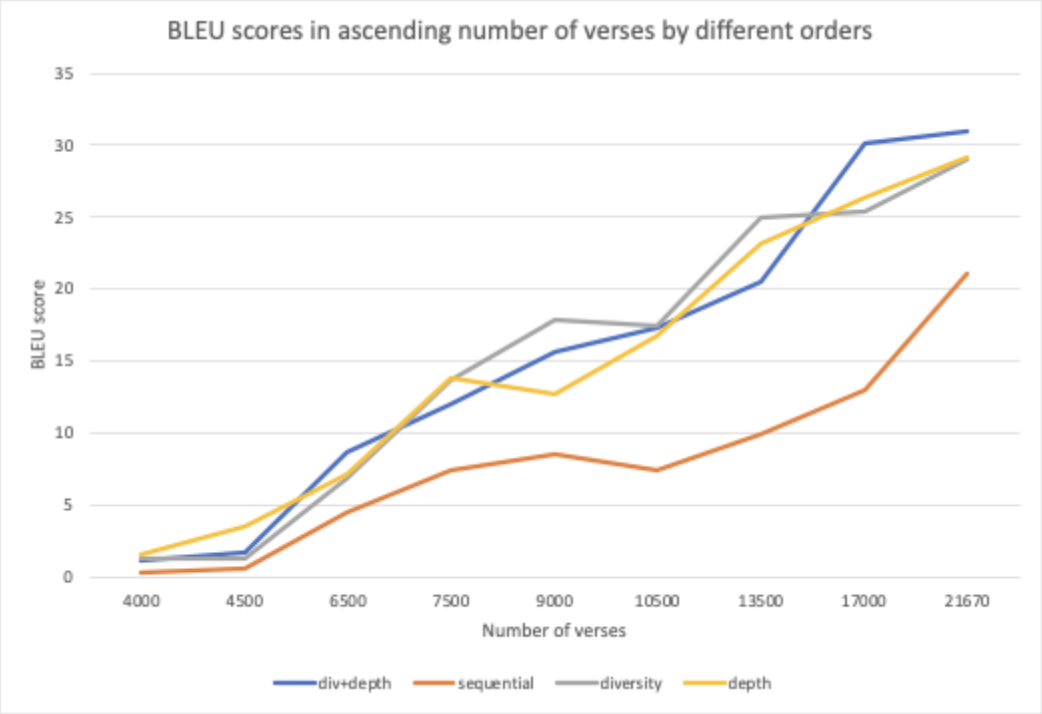

Figure 1: BLEU scores in ascending number of verses by different orders

Type of Order	Number of Verses								
	4000	4500	6500	7500	9000	10500	13500	17000	21670
div+depth	1.14	1.78	8.63	12.07	15.6	17.32	20.54	30.2	30.9
sequential	0.36	0.67	4.47	7.47	8.52	7.41	9.89	13.07	21.08
diversity	1.30	1.31	6.94	13.64	17.9	17.45	25	25.4	29.01
depth	1.52	3.51	7.17	13.83	12.77	16.73	23.18	26.3059	29.22

Table 8: Full BLEU Score Table

Comparison	Diversity + depth vs. sequential	Diversity vs. sequential	Depth vs. sequential
T-Value	4.07	4.25	4.53
P-Value	0.004	0.003	0.002

Table 9: 3 Paired T-Test

By improving the BLEU scores using a smaller number of verses, we believe that using strategic order based on diversity, depth or combination of both will contribute directly to reducing time and costs that are related to translation projects.

7. CONCLUSION

Finding translators to train enormous number of lingual corpora, deciding which texts to translate, paying for the translations, and dealing with the problem's computational complexity due to the massive amount of training data all add up to a lot of money, resources, and man hours for our client. The Joey NMT model that we have created will aid not only in reducing the cost of hiring translators but also in reducing the computation cost. Furthermore, similar accuracy would now be achieved with relatively less data and will therefore result in less processing time, leading to a reduction in computational complexity. In our study, we have assumed that the translations done by the language experts have almost perfect accuracy. We strongly believe that including external supplemental data will further improve the accuracy rate. That data, on the other hand, should have a comparable level of diversity and depth when compared to Bible text.

REFERENCES

[1] Martinus, L., & Abbott, J. Z. (2019). A focus on neural machine translation for african languages. arXiv preprint arXiv:1906.05685.

[2] Van Der Wees, M., Bisazza, A., & Monz, C. (2017). Dynamic data selection for neural machine translation. arXiv preprint arXiv:1708.00712.

[3] Ranathunga, S., Lee, E. S. A., Skenduli, M. P., Shekhar, R., Alam, M., & Kaur, R. (2021). Neural machine translation for low-resource languages: A survey. arXiv preprint arXiv:2106.15115.

[4] Xinyi Wang and Graham Neubig (2019). Optimizing Data Selection for Multilingual Neural Machine Translation arXiv:1905.08212

[5] Liwei Wu, Shanbo Cheng, Mingxuan Wang, Lei Li (2021) Language Tags Matter for Zero-Shot Neural Machine Translation arXiv:2106.07930v1