Active Learning for Massively Parallel Translation of Constrained Text into Low Resource Languages

Zhong Zhou zhongzhou@cmu.edu
Alex Waibel alex@waibel.com
Language Technology Institute, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh PA 15213

Abstract

We translate a closed text that is known in advance and available in many languages into a new and severely low resource language. Most human translation efforts adopt a portion-based approach to translate consecutive pages/chapters in order, which may not suit machine translation. We compare the portion-based approach that optimizes coherence of the text locally with the random sampling approach that increases coverage of the text globally. Our results show that the random sampling approach performs better. When training on a seed corpus of ∼1,000 lines from the Bible and testing on the rest of the Bible (∼30,000 lines), random sampling gives a performance gain of +11.0 BLEU using English as a simulated low resource language, and +4.9 BLEU using Eastern Pokomchi, a Mayan language. Furthermore, we compare three ways of updating machine translation models with increasing amount of human post-edited data through iterations. We find that adding newly post-edited data to training after vocabulary update without self-supervision performs the best. We propose an algorithm for human and machine to work together seamlessly to translate a closed text into a severely low resource language.

1 Introduction

Machine translation has flourished ever since the first computer was made (Hirschberg and Manning, 2015; Popel et al., 2020). Over the years, human translation is assisted by machine translation to remove human bias and translation capacity limitations (Koehn and Haddow, 2009; Li et al., 2014; Savoldi et al., 2021; Bowker, 2002; Bowker and Fisher, 2010; Koehn, 2009). By learning human translation taxonomy and post-editing styles, machine translation borrows many ideas from human translation to improve performance through active learning (Settles, 2012; Carl et al., 2011; Denkowski, 2015). We propose a workflow to bring human translation and machine translation to work together seamlessly in translation of a closed text into a severely low resource language as shown in Figure 1 and Algorithm 1.

Given a closed text that has many existing translations in different languages, we are interested in translating it into a severely low resource language well. Researchers recently have shown achievements in translation using very small seed parallel corpora in low resource languages (Lin et al., 2020; Qi et al., 2018; Zhou et al., 2018a). Construction methods of such seed corpora are therefore pivotal in translation performance. Historically, this is mostly determined by field linguists' experiential and intuitive discretion. Many human translators employ a portion-based strategy when translating large texts. For example, translation of the book "The Little Prince" may be divided into smaller tasks of translating 27 chapters, or even smaller translation units like a few consecutive pages. Each translation unit contains consecutive

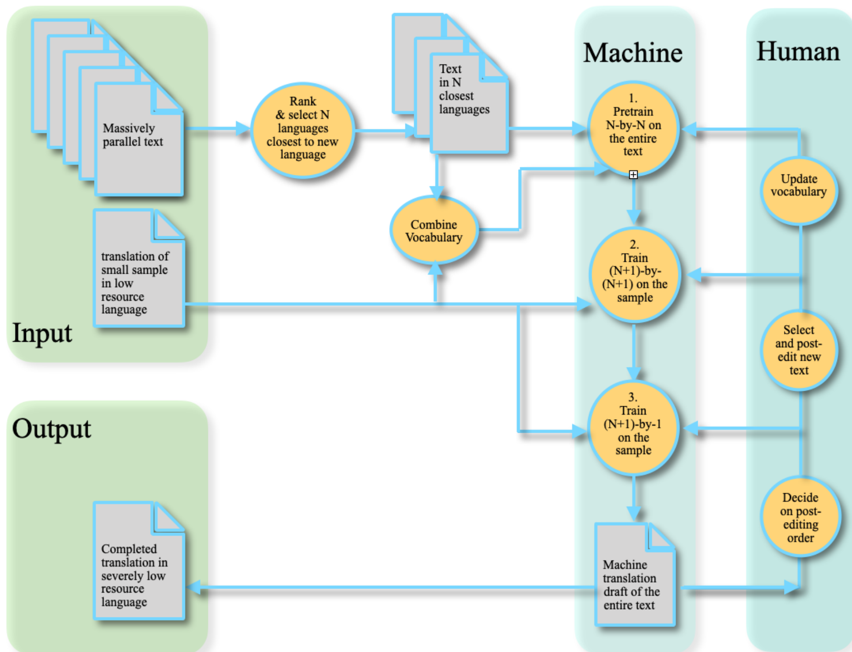


Figure 1: Proposed joint human machine translation sequence for a given closed text.

sentences. Consequently, machine translation often uses seed corpora that are chosen based on human translators' preferences, but may not be optimal for machine translation.

We propose to use a random sampling approach to build seed corpora when resources are extremely limited. In other words, when field linguists have limited time and resources, which lines would be given priority? Given a closed text, we propose that it would beneficial if field linguists translate randomly sampled ∼1,000 lines first, getting the first machine translated draft of the whole text, and then post-edit to obtain final translation of each portion iteratively as shown in Algorithm 1. We recognize that the portion-based translation is very helpful in producing quality translation with formality, cohesion and contextual relevance. Thus, our proposed way is not to replace the portion-based approach, but instead, to get the best of both worlds and to expedite the translation process as shown in Figure 1.

The main difference of the two approaches is that the portion-based approach focuses on preserving coherence of the text locally, while the random-sampling approach focuses on increasing coverage of the text globally. Our results show that the random sampling approach performs better. When training on a seed corpus of ∼1,000 lines from the Bible and testing on the rest of the Bible (∼30,000 lines), random sampling beats the portion-based approach by +11.0 BLEU using English as a simulated low resource language on a family of languages ranked by distortion, and by +4.9 using a Mayan language, Eastern Pokomchi, training on a family of languages based on linguistic definition. Using random sampling, machine translation is able to produce an apt first draft of the whole text that expedites the subsequent translation iterations.

Moreover, we compare three different ways of incorporating incremental post-edited data during the translation process. We find that self-supervision using the whole translation draft affects performance adversely, and is best to be avoided. We also show that adding the newly post-edited text to training with vocabulary update performs the best.

Algorithm 1: Proposed joint human machine translation sequence for a given closed text.

Input: A text of N lines consisting multiple books/portions, parallel in L source languages
Output: A full translation in the target low resource language, l'
0. Initialize translation size, $n = 0$, vocabulary size, $v = 0$, vocabulary update size, $\triangle v = 0$;
1. Randomly sample S ($\sim$1,000) sentences with vocabulary size v_S for human translators to produce the seed corpus, update $n = S$, $v = v_S$;
2. Rank and pick a family of close-by languages by linguistic, distortion or performance metric ;
while $n < N$ **do**
  if $\triangle v > 0$ **then**
    3. Pretrain on the full texts of neighboring languages ;
  4. Train on the n sentences of all languages in multi-source multi-target configuration ;
  5. Train on the n sentences of all languages in multi-source single-target configuration ;
  6. Combine translations from all source languages using the centeredness measure ;
  7. Review all books/portions of the translation draft ;
  8. Pick a book/portion with n' lines and v' more vocabulary ;
  9. Complete human post-editing of the portion chosen, $v = v + v'$, $n = n + n'$, $\triangle v = v'$;
return full translation co-produced by human (Step 1, 7-9) and machine (Step 0, 2-6) translation ;

2 Related Works

2.1 Human Translation and Machine Translation

Machine translation began about the same time as the first computer (Hirschberg and Manning, 2015; Popel et al., 2020). Over the years, human translators have different reactions to machine translation advances, mixed with doubt or fear (Hutchins, 2001). Some researchers study human translation taxonomy for machine to better assist human translation and post-editing efforts (Carl et al., 2011; Denkowski, 2015). Human translators benefit from machine assistance as human individual bias and translation capacity limitations are compensated for by large-scale machine translation (Koehn and Haddow, 2009; Li et al., 2014; Savoldi et al., 2021; Bowker, 2002; Bowker and Fisher, 2010; Koehn, 2009). On the other hand, machine translation benefits from professional human translators' context-relevant and culturally-appropriate translation and post-editing efforts (Hutchins, 2001). Severely low resource translation is a fitting ground for close human machine collaboration (Zong, 2018; Carl et al., 2011; Martínez, 2003).

2.2 Severely Low Resource Text-based Translation

Many use multiple rich-resource languages to translate to a low resource language using multilingual methods (Johnson et al., 2017; Ha et al., 2016; Firat et al., 2016; Zoph and Knight, 2016; Zoph et al., 2016; Adams et al., 2017; Gillick et al., 2016; Zhou et al., 2018a,b). Some use data selection for active learning (Eck et al., 2005). Some use as few as $\sim$4,000 lines (Lin et al., 2020; Qi et al., 2018) and $\sim$1,000 lines (Zhou and Waibel, 2021) of data. Some do not use low resource data (Neubig and Hu, 2018; Karakanta et al., 2018).

2.3 Active Learning and Random Sampling

Active learning has long been used in machine translation (Settles, 2012; Ambati, 2012; Eck et al., 2005; Haffari and Sarkar, 2009; González-Rubio et al., 2012; Miura et al., 2016; Gangadharaiah et al., 2009). Random sampling and data selection has been successful (Kendall and Smith, 1938; Knuth, 1991; Clarkson and Shor, 1989; Sennrich et al., 2015; Hoang et al., 2018; He et al., 2016; Gu et al., 2018). The mathematician Donald Knuth uses the population of Menlo Park to illustrate the value of random sampling (Knuth, 1991).

Book	Author	Books	Chapters	Pages	Languages
The Bible	Multiple	66	1,189	1,281	689
The Little Prince	Antoine de Saint Exupéry	1	27	96	382
Dao De Jing	Laozi	1	81	$\sim$10	>250
COVID-19 Wiki Page	Multiple	1	1	$\sim$50	155
The Alchemist	Paulo Coelho	1	2	163	70
Harry Potter	J. K. Rowling	7	199	3,407	60
The Lord of the Rings	J. R. R. Tolkien	6	62	1,037	57
Frozen Movie Script	Jennifer Lee	1	112	$\sim$40	41
The Hand Washing Song	Multiple	1	1	1	28
Dream of the Red Chamber	Xueqin Cao	2	120	2500	23
Les Misérables	Victor Hugo	68	365	1,462	21

Table 1: Examples of different texts with the number of languages translated to date (UNESCO, 1932; Mayer and Cysouw, 2014; de Saint-Exupéry, 2019; Laozi, 2019; Fung et al., 2020; Coelho, 2015; Rowling, 2019; Tolkien, 2012; Lee, 2013; Thampi et al., 2020; Xueqin, 2016; Hugo, 1863).

3 Methodology

We train our models using a state-of-the-art multilingual transformer by adding language labels to each source sentence (Johnson et al., 2017; Ha et al., 2016; Zhou et al., 2018a,b). We borrow the order-preserving named entity translation method by replacing each named entity with `__NE`s (Zhou et al., 2018b) using a multilingual lexicon table that covers 124 source languages and 2,939 named entities (Zhou and Waibel, 2021). For example, the sentence "Somchai calls Juan" is transformed to "`__opt_src_en __opt_tgt_ca __NE0` calls `__NE1`" to translate to Chuj. We use families of close-by languages constructed by ranking 124 source languages by distortion measure (*FAMD*), performance measure (*FAMP*) and linguistic family (*FAMO*$^{+}$); the distortion measure ranks languages by decreasing probability of zero distortion, while the performance measure incorporates an additional probability of fertility equalling one (Zhou and Waibel, 2021). Using families constructed, we pretrain our model first on the whole text of nearby languages, then we train on the $\sim$1,000 lines of low resource data and the corresponding lines in other languages in a multi-source multi-target fashion. We finally train on the $\sim$1,000 lines in a multi-source single-target fashion (Zhou and Waibel, 2021).

We combine translations of all source languages into one. Let all N translations be $t_i, i = 1, \ldots, N$ and let similarity between translations t_i and t_j be S_{ij}. We rank all translations according to how centered it is with respect to other sentences by summing all its similarities to the rest through $\sum_j S_{ij}$ for $i = 1, \ldots, N$. We take the most centered translation for every sentence, $\max_i \sum_j S_{ij}$, to build the combined translation output. The expectation of the combined score is higher than that of any of the source languages (Zhou and Waibel, 2021).

Our work differs from the past research in that we put low resource translation into the broad collaborative scheme of human machine translation. We compare the portion-based approach with the random sampling approach in building seed corpora. We also compare three methods of updating models with increasing amount of human post-edited data. We add the newly post-edited data to training in three ways: with vocabulary update, without vocabulary update, or incorporating the whole translation draft in a self-supervised fashion additionally. For best performance, we build the seed corpus by random sampling, update vocabulary iteratively, and add newly post-edited data to training without self-supervision. We also have a larger test set, we test on $\sim$30,000 lines rather than $\sim$678 lines from existing research.

We propose a joint human machine translation workflow in Algorithm 1. After pretraining

Input Language Family														
By Linguistics					By Distortion					By Performance				
FAMO$^+$					*FAMD*					*FAMP*				
Training	*Luke*		*Rand*		Training	*Luke*		*Rand*		Training	*Luke*		*Rand*	
Testing	*Best*	*All*	*Best*	*All*	Testing	*Best*	*All*	*Best*	*All*	Testing	*Best*	*All*	*Best*	*All*
Combined	38.2	21.9	47.7	31.3	Combined	38.4	22.9	49.6	33.9	Combined	40.3	23.7	48.8	33.2
German	35.8	20.0	45.4	29.4	German	36.8	20.8	47.2	31.5	German	37.6	21.3	46.5	30.9
Danish	36.8	18.9	43.3	28.8	Danish	37.4	19.6	44.7	30.8	Danish	38.5	19.9	44.4	30.2
Dutch	36.2	20.3	45.3	29.9	Dutch	36.3	21.0	47.1	32.3	Dutch	37.8	21.6	46.3	31.6
Norwegian	36.6	20.2	45.1	29.7	Norwegian	36.9	20.9	46.5	31.7	Norwegian	37.6	21.2	46.1	31.2
Swedish	35.2	19.6	45.1	29.0	Afrikaans	38.4	22.2	48.0	33.1	Afrikaans	39.6	22.9	47.5	32.4
Spanish	36.8	21.6	45.1	30.3	Marshallese	35.3	21.6	47.1	31.5	Spanish	38.9	22.9	46.6	31.7
French	36.1	19.7	44.6	28.9	French	36.3	20.3	46.0	30.9	French	37.4	21.7	45.4	30.2
Italian	36.9	20.5	43.5	29.7	Italian	37.1	21.0	45.2	31.7	Italian	38.8	21.8	44.6	31.1
Portuguese	32.5	15.8	35.2	24.4	Portuguese	33.3	16.5	38.1	26.9	Portuguese	34.0	16.3	36.2	25.8
Romanian	34.9	19.3	43.0	28.8	Frisian	36.3	21.6	47.7	32.4	Frisian	38.0	22.3	47.4	31.8

Table 2: Performance training on 1,093 lines of Eastern Pokomchi data on *FAMO*$^+$, *FAMD* and *FAMP*. We train using the portion-based approach in *Luke*, and using random sampling in *Rand*. During testing, *Best* is the book with highest BLEU score, and *All* is the performance on ~30,000 lines of test data.

on neighboring languages in Step 3, we iteratively train on the randomly sampled seed corpus of low resource data in Step 4 and 5. The reason we include both Step 4 and 5 in our algorithm is because training both steps iteratively performs better than training either one (Zhou and Waibel, 2021). Our model produces a translation draft of the whole text. Since the portion-based approach has the advantage with formality, cohesion and contextual relevance, human translators may pick and post-edit portion-by-portion iteratively. The newly post-edited data with updated vocabulary is feed back to the machine translation models without self-supervision. In this way, machine translation systems rely on quality parallel corpora that are incrementally produced by human translators. Human translators lean on machine translation for quality translation draft to expedite translation. This creates a synergistic collaboration between human and machine.

4 Data

We work on the Bible in 124 source languages (Mayer and Cysouw, 2014), and have experiments for English, a simulated language, and Eastern Pokomchi, a Mayan language. We train on ~1,000 lines of low resource data and on full texts for all the other languages. We aim to translate the rest of the text (~30,000 lines) into the low resource language. In pretraining, we use 80%, 10%, 10% split for training, validation and testing. In training, we use 3.3%, 0.2%, 96.5% split for training, validation and testing. Our test size is >29 times of the training size. We use the book "Luke" for the portion-based approach as suggested by many human translators.

Training on ~100 million parameters with Geforce RTX 2080 Ti, we employ a 6-layer encoder and a 6-layer decoder with 512 hidden states, 8 attention heads, 512 word vector size, 2,048 hidden units, 6,000 batch size, 0.1 label smoothing, 2.5 learning rate, 0.1 dropout and attention dropout, an early stopping patience of 5 after 190,000 steps, “BLEU” validation metric, “adam” optimizer and “noam” decay method (Klein et al., 2017; Papineni et al., 2002). We increase patience to 25 for larger data in the second stage of training in Figure 2a and 2b.

Input Language Family														
By Linguistics					By Distortion					By Performance				
FAMO$^+$					*FAMD*					*FAMP*				
Training	*Luke*		*Rand*		Training	*Luke*		*Rand*		Training	*Luke*		*Rand*	
Testing	*Best*	*All*	*Best*	*All*	Testing	*Best*	*All*	*Best*	*All*	Testing	*Best*	*All*	*Best*	*All*
Combined	23.1	8.6	24.4	13.5	Combined	23.2	8.5	22.7	12.6	Combined	22.2	7.2	20.3	10.9
Chuj	21.8	8.0	21.3	12.8	Chuj	21.9	8.5	20.2	12.0	Chuj	21.8	7.2	18.0	10.3
Cakchiquel	22.2	7.9	22.4	13.0	Cakchiquel	22.3	7.9	21.8	12.2	Cakchiquel	21.2	6.9	19.1	10.5
Guajajara	19.7	7.0	18.8	11.8	Guajajara	19.1	6.9	18.0	11.2	Guajajara	18.8	5.9	15.1	9.5
Mam	22.2	8.6	24.1	13.7	Russian	22.2	7.3	17.4	11.8	Mam	21.7	7.5	21.4	11.1
Kanjobal	21.8	8.1	22.3	13.1	Toba	21.9	8.3	21.8	12.5	Kanjobal	21.5	7.1	18.7	10.6
Cuzco	22.3	7.8	22.5	12.9	Myanmar	19.1	5.3	13.3	9.8	Thai	21.8	6.3	15.7	10.2
Ayacucho	21.6	7.6	23.3	12.8	Slovenský	22.1	7.5	18.5	12.0	Dadibi	19.9	6.2	17.8	9.8
Bolivian	22.2	7.8	22.3	12.9	Latin	21.9	7.8	20.4	12.2	Gumatj	19.1	3.8	11.7	4.7
Huallaga	22.2	7.7	22.7	12.8	Ilokano	22.6	8.4	22.4	12.5	Navajo	21.3	6.5	17.4	10.5
Aymara	21.4	7.5	23.0	12.7	Norwegian	22.6	8.3	22.0	12.6	Kim	21.6	7.0	17.5	10.7

Table 3: Performance training on 1,086 lines of Eastern Pokomchi data on *FAMO*$^+$, *FAMD* and *FAMP*. We train using the portion-based approach in *Luke*, and using random sampling in *Rand*. During testing, *Best* is the book with highest BLEU score, and *All* is the performance on ∼30,000 lines of test data.

5 Results

We observe that random sampling performs better than the portion-based approach. Random sampling gives a performance gain of +11.0 for English on FAMD and +4.9 for Eastern Pokomchi on FAMO$^+$ in Table 2 and 3. The performance gain for Eastern Pokomchi may be lower because Mayan languages are morphologically rich, complex, isolated and opaque (Aissen et al., 2017; Clemens et al., 2015; England, 2011). English is closely related to many languages due to colonization and globalization even though it is artificially constrained in size (Bird, 2020). This may explain why Eastern Pokomchi benefits less.

To simulate human translation efforts in Step 7 and 8 in Algorithm 1, we rank 66 books of the Bible by BLEU score on English's FAMD and Eastern Pokomchi's FAMO$^+$. We assume that BLEU ranking is available to us to simulate human judgment. In reality, this step is realized by human translators skimming through the translation draft and comparing performances of different books by intuition and experience. In Section 6, we will discuss the limitation of this assumption. Performance ranking of the simulated low resource language may differ from that of the actual low resource language. But the top few may coincide because of the nature of the text, independent of the language. In our results, we observe that the narrative books performs better than the philosophical or poetic books. The book "1 Chronicles" performs best for both English and Eastern Pokomchi, and the book "Philemon" performs worst for both languages. A possible explanation is that "1 Chronicles" is mainly narrative, and contains many named entities that are translated well by the order-preserving lexiconized model. If we compare BLEU scores of the best-performing book, random sampling outperforms the portion-based approach by +11.2 on English's FAMD, and by +1.3 on Eastern Pokomchi's FAMO$^+$.

In Table 4, we compare three different ways of updating the machine translation models by adding a newly post-edited book that human translators produced. We call the baseline without addition of the new book *Seed*. *Updated-Vocab* adds the new book to training with updated vocabulary while *Old-Vocab* skips the vocabulary update. *Self-Supervised* adds the whole translation draft of ∼30,000 lines to pretraining in addition to the new book. Self-supervision

Source	*Seed*	*Self-Supervised*	*Old-Vocab*	*Updated-Vocab*
Combined	33.9	29.4 (-4.5)	36.3 (+2.4)	36.7 (+2.8)
Danish	31.5	26.8 (-4.7)	33.1 (+1.6)	33.7 (+2.2)
Norwegian	30.8	27.6 (-3.2)	34.1 (+3.3)	34.7 (+3.9)
Italian	32.3	27.3 (-5.0)	34.1 (+1.8)	34.6 (+2.3)
Afrikaans	31.7	28.8 (-2.9)	35.6 (+3.9)	36.0 (+4.3)
Dutch	33.1	28.0 (-5.1)	34.6 (+1.5)	35.1 (+2.0)
Portuguese	31.5	23.6 (-7.9)	29.1 (-2.4)	29.8 (-0.7)
French	30.9	26.8 (-4.1)	33.3 (+2.4)	33.9 (+3.0)
German	31.7	27.4 (-4.3)	33.8 (+2.1)	34.4 (+2.7)
Marshallese	26.9	27.5 (+0.6)	33.8 (+6.9)	34.4 (+7.5)
Frisian	32.4	28.2 (-4.2)	34.7 (+2.3)	35.3 (+2.9)

Table 4: Comparing three ways of adding the newly post-edited book “1 Chronicles”. *Seed* is the baseline of training on the seed corpus alone, *Old-Vocab* skips the vocabulary update while *Updated-Vocab* has vocabulary update. *Self-Supervised* adds the complete translation draft in addition to the new book.

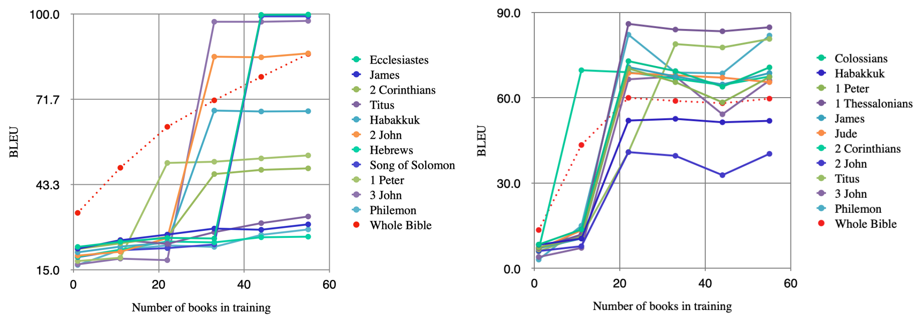
refers to using the small seed corpus to translate the rest of the text which is subsequently used to train the model. We observe that the *Self-Supervised* performs the worst among the three. Indeed, *Self-Supervised* performs even worse than the baseline *Seed*. This shows that quality is much more important than quantity in severely low resource translation. It is better for us not to add the whole translation draft to the pretraining as it affects performance adversely.

On the other hand, we see that both *Updated-Vocab* and *Old-Vocab* performs better than *Seed* and *Self-Supervised*. *Updated-Vocab*’s performance is better than *Old-Vocab*. An explanation could be that *Updated-Vocab* has more expressive power with updated vocabulary. Therefore, in our proposed algorithm, we prefers vocabulary update in each iteration. If the vocabulary has not increased, we may skip pretraining to expedite the process.

We show how the algorithm is put into practice for English and Eastern Pokomchi in Figure 2a and 2b. We take the worst-performing 11 books as the held-out test set, and divide the other 55 books of the Bible into 5 portions. Each portion contains 11 books. We translate the text by using the randomly sampled $\sim$1,000 lines of seed corpus first, and then proceed with human machine translation in Algorithm 1 in 5 iterations with increasing number of post-edited portions. The red dotted line is the overall performance of the whole text excluding the seed corpus. We observe that the red dotted curve is steadily increasing for both languages. However, since we are interested in the test results of the held-out set, we evaluate only on the solid lines plotted.

For English, we observe that philosophical books like “Ecclesiastes” and poetry books like “Song of Solomon” perform very badly in the beginning, but begin to achieve above 90 BLEU scores after adding 33 books of training data. The high performance is due to the multilingual cross-lingual transfer and this is the main reason why we set up our problem as translation of a closed text that are available in many languages to the low resource language. However, some books like “Philemon”, “Hebrews”, “James”, “Titus” remains difficult to translate even after adding 55 books of training data. This shows that adding data may benefit some books more than the others. A possible explanation is that there are multiple authors of the Bible, and books differ from each other in style and content. Some books are closely related to each other, and may benefit from translations of other books. But some may be very different and benefit much less.

For Eastern Pokomchi, even though the performance of the most difficult 11 books never reach the near perfect BLEU score of 90s like that of English experiments, all books has BLEU



(a) English

(b) Eastern Pokomchi

Figure 2: Performance of the most difficult 11 books with increasing number of training books.

scores that are steadily increasing. Surprisingly, we observe good performance with the books that remain difficult with large training data in the English experiments. “Philemon”, for example, increases to a BLEU score of 81.9 with 55 books of training data in Eastern Pokomchi while it has a BLEU score of 28.4 with 55 books of training data in English. This surprising result shows that what is difficult for simulated low resource languages may not be as difficult for real low resource languages. Even though Eastern Pokomchi gives a lower overall BLEU score than English, it has a better generalization to the most difficult book.

6 Conclusion

We propose to use random sampling to build seed parallel corpora instead of using the portion-based approach in severely low resource settings. Training on $\sim$1,000 lines, the random sampling approach outperforms the portion-based approach by +11.0 for English’s FAMD, and by +4.9 for Eastern Pokomchi’s FAMO^{+}. We also compare three different ways of updating the machine translation models by adding newly post-edited data iteratively. We find that vocabulary update is necessary, but self-supervision by pretraining with whole translation draft is best to be avoided.

One limitation of our work is that in real life scenarios, we do not have the reference text in low resource languages to produce the BLEU scores to decide the post-editing order. Consequently, field linguists need to skim through and decide the post-editing order based on intuition. However, computational models can still help. One potential way to tackle it is that we can train on $\sim$1,000 lines from another language with available text and test on the 66 books. Since our results show that the literary genre plays important role in the performance ranking, it would be reasonable to determine the order using a “held-out language” and then using that to determine order in the target low resource language. In the future, we would like to work with human translators who understand and speak low resource languages.

Another concern human translators may have is the creation of randomly sampled seed corpora. To gauge the amount of interest or inertia, we have interviewed some human translators and many are interested. However, it is unclear whether human translation quality of randomly sampled data differs from that of the traditional portion-based approach. We hope to work with human translators closely to determine whether the translation quality difference is manageable.

We are also curious how our model will perform with large literary works like “Lord of the Rings” and "Les Misérables". We would like to see whether it will translate well with philosophical depth and literary complexity. However, these books often have copyright issues and are not as easily available as the Bible data. We are interested in collaboration with teams who have multilingual data for large texts, especially multilingual COVID-19 data.

References

Adams, O., Makarucha, A., Neubig, G., Bird, S., and Cohn, T. (2017). Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 937–947.

Aissen, J., England, N. C., and Maldonado, R. Z. (2017). *The Mayan languages*. Taylor & Francis.

Ambati, V. (2012). *Active learning and crowdsourcing for machine translation in low resource scenarios*. PhD thesis, Carnegie Mellon University.

Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519.

Bowker, L. (2002). *Computer-aided translation technology: A practical introduction*. University of Ottawa Press.

Bowker, L. and Fisher, D. (2010). Computer-aided translation. *Handbook of translation studies*, 1:60–65.

Carl, M., Dragsted, B., and Jakobsen, A. L. (2011). A taxonomy of human translation styles. *Translation journal*, 16(2):155–168.

Clarkson, K. L. and Shor, P. W. (1989). Applications of random sampling in computational geometry, ii. *Discrete & Computational Geometry*, 4(5):387–421.

Clemens, L. E., Coon, J., Pedro, P. M., Morgan, A. M., Polinsky, M., Tandet, G., and Wagers, M. (2015). Ergativity and the complexity of extraction: A view from mayan. *Natural Language & Linguistic Theory*, 33(2):417–467.

Coelho, P. (2015). *The alchemist*. HarperOne; 25th edition.

de Saint-Exupéry, A. (2019). *El Principito: The Little Prince*. Editorial Verbum.

Denkowski, M. (2015). Machine translation for human translators. *Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania*.

Eck, M., Vogel, S., and Waibel, A. (2005). Low cost portability for statistical machine translation based on n-gram frequency and tf-idf. In *International Workshop on Spoken Language Translation*.

England, N. C. (2011). *A grammar of Mam, a Mayan language*. University of Texas Press.

Firat, O., Cho, K., and Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 866–875.

Fung, T. L. et al. (2020). COVID-19. `https://en.wikipedia.org/wiki/COVID-19`. [Online; accessed 24-May-2021].

Gangadharaiah, R., Brown, R. D., and Carbonell, J. G. (2009). Active learning in example-based machine translation. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 227–230.

Gillick, D., Brunk, C., Vinyals, O., and Subramanya, A. (2016). Multilingual language processing from bytes. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 1296–1306.

González-Rubio, J., Ortiz-Martínez, D., and Casacuberta, F. (2012). Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 245–254. Association for Computational Linguistics.

Gu, J., Wang, Y., Chen, Y., Cho, K., and Li, V. O. (2018). Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*.

Ha, T.-L., Niehues, J., and Waibel, A. (2016). Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Haffari, G. and Sarkar, A. (2009). Active learning for multilingual statistical machine translation. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.-Y., and Ma, W.-Y. (2016). Dual learning for machine translation. *Advances in neural information processing systems*, 29:820–828.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Hoang, V. C. D., Koehn, P., Haffari, G., and Cohn, T. (2018). Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

Hugo, V. (1863). *Les Miserables*. C. Lassalle.

Hutchins, J. (2001). Machine translation and human translation: in competition or in complementation. *International Journal of Translation*, 13(1-2):5–20.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Karakanta, A., Dehdari, J., and van Genabith, J. (2018). Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1):167–189.

Kendall, M. G. and Smith, B. B. (1938). Randomness and random sampling numbers. *Journal of the royal Statistical Society*, 101(1):147–166.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). Opennmt: Open-source toolkit for neural machine translation. *Proceedings of the 55th annual meeting of the Association for Computational Linguistics, System Demonstrations*, pages 67–72.

Knuth, D. E. (1991). *3: 16 Bible texts illuminated*. AR Editions, Inc.

Koehn, P. (2009). A process study of computer-aided translation. *Machine Translation*, 23(4):241–263.

Koehn, P. and Haddow, B. (2009). Interactive assistance to human translators using statistical machine translation methods. *MT Summit XII*.

Laozi (2019). *Dao de jing*. University of California Press.

Lee, J. (2013). Frozen.

Li, H., Graesser, A. C., and Cai, Z. (2014). Comparison of google translation with human translation. In *The Twenty-Seventh International Flairs Conference*.

Lin, Z., Pan, X., Wang, M., Qiu, X., Feng, J., Zhou, H., and Li, L. (2020). Pre-training multilingual neural machine translation by leveraging alignment information. *arXiv preprint arXiv:2010.03142*.

Martínez, L. G. (2003). *Human Translation Versuv Machine Translation and Full Post-editing of Raw Machine Translation Output*. Citeseer.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel bible corpus. *Oceania*, 135(273):40.

Miura, A., Neubig, G., Paul, M., and Nakamura, S. (2016). Selecting syntactic, non-redundant segments in active learning for machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–29.

Neubig, G. and Hu, J. (2018). Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., and Žabokrtskỳ, Z. (2020). Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Qi, Y., Sachan, D. S., Felix, M., Padmanabhan, S. J., and Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation? *arXiv preprint arXiv:1804.06323*.

Rowling, J. (2019). Harry potter. *The 100 Greatest Literary Characters*, page 183.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., and Turchi, M. (2021). Gender bias in machine translation. *arXiv preprint arXiv:2104.06001*.

Sennrich, R., Haddow, B., and Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Settles, B. (2012). Active learning. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):1–114.

Thampi, N., Longtin, Y., Peters, A., Pittet, D., and Overy, K. (2020). It's in our hands: a rapid, international initiative to translate a hand hygiene song during the covid-19 pandemic. *Journal of Hospital Infection*, 105(3):574–576.

Tolkien, J. R. R. (2012). *The Lord of the Rings: One Volume*. Houghton Mifflin Harcourt.

UNESCO, I. T. (1932). World bibliography of translation.

Xueqin, C. (2016). *Dream of the Red Chamber*. Editorial Axioma.

Zhou, Z., Sperber, M., and Waibel, A. (2018a). Massively parallel cross-lingual learning in low-resource target language translation. In *Proceedings of the 3rd conference on Machine Translation Worshop of the 23rd Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Zhou, Z., Sperber, M., and Waibel, A. (2018b). Paraphrases as foreign languages in multilingual neural machine translation. *Proceedings of the Student Research Workshop at the 56th Annual Meeting of the Association for Computational Linguistics*.

Zhou, Z. and Waibel, A. (2021). Family of origin and family of choice: Massively parallel lexiconized iterative pretraining for severely low resource text-based translation. *Proceedings of the 3rd Workshop on Research in Computational Typology and Multilingual NLP of the 20th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*.

Zong, Z. (2018). Research on the relations between machine translation and human translation. In *Journal of Physics: Conference Series*, page 062046. IOP Publishing.

Zoph, B. and Knight, K. (2016). Multi-source neural translation. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 30–34.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 21st Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.