Family of Origin and Family of Choice: Massively Parallel Lexiconized Iterative Pretraining for Severely Low Resource Machine Translation

Zhong Zhou
Carnegie Mellon University
zhongzhou@cmu.edu

Alex Waibel
Carnegie Mellon University
alex@waibel.com

Abstract

We translate a closed text that is known in advance into a severely low resource language by leveraging massive source parallelism. Our contribution is four-fold. Firstly, we rank 124 source languages empirically to determine their closeness to the low resource language and select the top few. We call the linguistic definition of language family Family of Origin (FAMO), and we call the empirical definition of higher-ranked languages using our metrics Family of Choice (FAMC). Secondly, we build an Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer (IPML) to train on $\sim$1,000 lines of low-resource data from the Bible dataset and the medical EMEA dataset. Using English as a hypothetical low-resource language to translate from Spanish, we obtain a +24.7 BLEU increase over a multilingual baseline, and a +10.2 BLEU increase over our asymmetric baseline. Thirdly, we also use a real severely low-resource Mayan language, Eastern Pokomchi. Finally, we add an order-preserving lexiconized component to translate named entities accurately. We build a massive lexicon table for 2,939 Bible named entities in 124 source languages, and include many that occur once and covers more than 66 severely low-resource languages. Training on randomly sampled 1,093 lines of low-resource data, we reach a 30.3 BLEU score for Spanish-English translation testing on 30,022 lines of Bible, and a 42.8 BLEU score for Portuguese-English translation on the medical EMEA dataset.

1 Introduction

We translate a closed text that is known in advance into a severely low resource language by leveraging massive source parallelism. In other words, we aim to translate well under three constraints: having severely small training data in the new target low resource language, having massive source language parallelism, having the same closed text across all languages. Generalization to other texts is preferable but not necessary in the goal of producing high quality translation of the closed text.

2020 is the year that we started the life-saving hand washing practice globally. Applications like translating water, sanitation, and hygiene (WASH) guidelines into severely low resource languages are very impactful in tribes like those in Papua New Guinea with 839 living languages (Gordon Jr, 2005; Simons and Fennig, 2017). Translating humanitarian texts like WASH guidelines with scarce data and expert help is key (Bird, 2020).

We focus on four challenges that are not addressed previously. Most multilingual transformer works that translate into low resource language limit their training data to available data in the same or close-by language families or the researchers’ intuitive discretion; and are mostly limited to less than 30 languages (Gu et al., 2018; Zhou et al., 2018a; Zhu et al., 2020). Instead, we examine ways to pick useful source languages from 124 source languages in a principled fashion. Secondly, most works require at least 4,000 lines of low resource data (Lin et al., 2020; Qi et al., 2018; Zhou et al., 2018a); we train on only $\sim$1,000 lines of low re-

| Eastern Pokomchi | English |
|-----------------|--------|
| **FAMD** | **FAMP** | **FAMD** | **FAMP** |
| Chuj* | Dadibi | Danish* | Dutch* |
| Cakchiquel* | Thai | Norwegian* | Afrikaans* |
| Guajajara* | Gumatj | Italian | Norwegian* |
| Toba | Navajo | Afrikaans* | German* |
| Myanmar | Cakchiquel* | Dutch* | Danish* |
| Slovenský | Kanjobal | Portuguese | Spanish |
| Latin | Guajajara* | French | Frisian* |
| Ilokano | Mam* | German* | Italian |
| Norwegian | Kim | Marshallese | French |
| Russian | Chuj* | Frisian* | Portuguese |

Table 1: Top ten languages closest to Eastern Pokomchi (left) and English (right) in ranking 124 source languages. FAMD and FAMP are two constructions of Family of Choice (FAMC) by distortion and performance metrics respectively. All are trained on $\sim$1,000 lines. We star those in Family of Origin.
source data. Thirdly, many works use rich resource languages as hypothetical low resource languages. In addition to English, our work includes Eastern Pokomchi, a Mayan language. Finally, most works do not treat named entities separately; we add an order-preserving lexiconized component for more accurate translation of named entities.

Our contribution is four-fold. Firstly, we rank the 124 source languages empirically to determine their closeness to the low resource language and choose the top few. We call the linguistic definition of language family Family of Origin (FAMO), and we call the empirical definition of higher-ranked languages using our metrics Family of Choice (FAMC). They often overlap, but may not coincide. We compare two different metrics of ranking linguistic distance to the low resource language.

Secondly, we build an Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer (IPML) training on \(\sim 1,000\) lines of low resource data. For training data, we distinguish between the complete graph (multi-source multi-target) configuration and the star graph (multi-source single-target) configuration. Using iterative pretraining, we get a +24.7 BLEU increase over a multilingual order-preserving lexiconized transformer baseline (MLc) in Spanish-English translation using English as a hypothetical low resource language, and a +10.2 BLEU increase over our asymmetric baseline. Training with the low resource language on both the source and target sides boosts translation into the target side. Training on randomly sampled 1,093 lines of low resource data, we reach a 30.3 BLEU score for Spanish-English translation testing on 30,022 lines of Bible. We have a 42.8 BLEU score for Portuguese-English translation on the medical EMEA dataset.

Thirdly, we use a real-life severely low resource Mayan language, Eastern Pokomchi, a Class 0 language (Joshi et al., 2020) as one of our experiment setups. In addition, we also use English as a hypothetical low resource language for easy evaluation.

Finally, we add an order-preserving lexiconized component to translate named entities accurately. To solve the variable-binding problem to distinguish “Ian calls Yi” from “Yi calls Ian” (Fodor and Pylyshyn, 1988; Graves et al., 2014; Zhou et al., 2018a), we build a massively parallel lexicon table for 2,939 Bible named entities for 124 source languages, and include many that only occur once and 66 severely low resource languages.

2 Related Works

2.1 Information Dissemination

Interactive Natural Language Processing (NLP) systems are classified into information assimilation, dissemination, and dialogue (Bird, 2020; Ranzato et al., 2015; Waibel and Fugen, 2008). Information assimilation involves information flow from low resource to rich resource language communities while information dissemination involves information flow from rich resource to low resource language communities. Taken together, they allow dialogue and interaction of different groups at eye level. Most work on information assimilation (Bérard et al., 2020; Earle et al., 2012; Brownstein et al., 2008). Few work on dissemination due to small data, less funding, few experts and limited writing system (Östling and Tiedemann, 2017; Zoph et al., 2016; Anastasopoulos et al., 2017; Adams et al., 2017; Bansal et al., 2017).

2.2 Machine Polyglotism and Pretraining

Recent research on machine polyglotism involves training machines to be adept in many languages by adding language labels in the training data with a single attention (Johnson et al., 2017; Ha et al., 2016; Firat et al., 2016; Gillick et al., 2016; Zhou et al., 2018b). Some explores data symmetry (Freitag and Firat, 2020; Birch et al., 2008; Lin et al., 2019). Zero-shot translation in severely low resource settings exploits the massive multilinguality, cross-lingual transfer, pretraining, iterative back-translation and freezing subnetworks (Lauscher et al., 2020; Nooralahzadeh et al., 2020; Wang et al., 2020; Li et al., 2020; Pfeiffer et al., 2020; Baziotis et al., 2020; Chronopoulou et al., 2020; Lin et al., 2020; Thompson et al., 2018; Luong et al., 2014; Wei et al., 2020; Dou et al., 2020).

2.3 Linguistic Distance

To construct linguistic distances (Hajič, 2000; Oncevay et al., 2020), researchers explore typological distance (Rama and Kolachina, 2012; Pienemann et al., 2005; Svalberg and Chuchu, 1998; Hansen et al., 2012; Comrie, 2005), lexical distance on the Swadesh list (Huang et al., 2007), normalized Levenshtein distance and Jaccard distance (Serva and Petroni, 2008; Holman et al., 2008; Adebara et al., 2020), sonority distance (Parker, 2012) and spectral distance (Dubossarsky et al., 2020).
3 Methodology

3.1 Multilingual Order-preserving Lexiconized Transformer

3.1.1 Multilingual Transformer

In training, each sentence is labeled with the source and target language label. For example, if we translate from Chuj (“ca”) to Cakchiquel (“ck”), each source sentence is tagged with __opt_src_ca __opt_tgt_ck. A sample source sentence is “__opt_src_ca __opt_tgt_ck Tec’b’ejec e b’a mach ex tzeyc’och Jehová yipoc e c’ool”.

We use a 6-layer encoder and a 6-layer decoder that are powered by 512 hidden states, 8 attention heads, 512 word vector size, a dropout of 0.1, an attention dropout of 0.1, 2,048 hidden transformer feed-forward units, a batch size of 6,000, “adam” optimizer, “noam” decay method, and a label smoothing of 0.1 and a learning rate of 2.5 on OpenNMT (Klein et al., 2017; Vaswani et al., 2017). We train on Geforce RTX 2080 Ti using ~100 million parameters for at least 190,000 steps, validate based on BLEU score and have early stopping patience of 5.

3.1.2 Star Versus Complete Configuration

We show two configurations of translation paths in Figure 1: star graph (multi-source single-target) configuration and complete graph (multi-source multi-target) configuration. The complete configuration data increases quadratically while the star configuration data increases linearly.

3.1.3 Order-preserving Lexiconized transformer

The variable binding problem issue is difficult in severely low resource scenario; most neural models cannot distinguish the subject and the object of a simple sentence like “Fatma asks her sister Wati to call Yi, the brother of Andika”, especially when all named entities appear once or never appear in training (Fodor and Pylyshyn, 1988; Graves et al., 2014). Recently, researchers use order-preserving lexiconized Neural Machine Translation models where named entities are sequentially tagged in a sentence as __NEs (Zhou et al., 2018a). The previous example becomes “__NE0 asks her sister __NE1 to call __NE2, the brother of __NE3”.

This method works under the assumption of translating a closed text known in advance, which matches our research goal. Its success relies on good coverage of named entities. To cover many named entities, we build on existing research literature (Wu et al., 2018; Zhou et al., 2018a) to construct a massively parallel lexicon table that covers 2,939 named entities across 124 languages in our Bible database. Our lexicon table is an expansion of the existing literature that covers 1,129 named entities (Wu et al., 2018). We add in 1,810 named entities that are in the extreme end of the tail occurring only once. We also include 66 more real-life severely low resource languages.

For every sentence pair, we build a target named entity decoding dictionary by using all target lexicons from the lexicon table that match with those in the source sentence. During the evaluation stage, we replace all the ordered __NEs using the target decoding dictionary to obtain our final translation.

Let us take the same example “Fatma asks her sister Wati to call Yi, the brother of Andika” and translate it to Chinese and German. Our tagged source sentence that translates to Chinese is “__opt_src_en __opt_tgt_zh __NE0 asks her mom __NE1 to call __NE2, the brother of __NE3”; and we use __opt_tgt_de for German. We create the source dictionary “__NE0: Fatma, __NE1: Wati, __NE2: Yi, __NE3: Andika” and the corresponding target dictionaries. With the dictionaries, we can successfully decode the transformer output to Chinese as “__NE0叫她的姐妹__NE1去打电话给__NE3的兄弟__NE2” and to German as “__NE0 bittet ihre Schwester __NE1 darum, __NE2, den Bruder __NE3, anzurufen”.

![Figure 1: (a) Complete graph configuration of translation paths (Many-to-many) in an example of multilingual translation. (b) Star configuration of translation paths (Many-to-one) using Indonesian as the low resource example.](image)
### 3.2 Ranking Source Languages

Many works on translation from multiple source languages into a single low resource language (Gu et al., 2018; Zhou et al., 2018a; Zhu et al., 2020). However, the languages in the training data are mostly limited to those within the same or close-by language families, or those with available data, or those chosen based on the researchers’ intuitive discretion. And they are mostly limited to less than 30 languages. Instead, we examine ways to pick useful source languages in a principled fashion motivated by cross-lingual impacts and similarities (Show-emark et al., 2016; Sapir, 1921; Odlin, 1989; Cenoz, 2001; Toral and Way, 2018; De Raad et al., 1997; Hermans, 2003; Specia et al., 2016). We find that using many languages that are distant to the target low resource language may produce marginal improvements, if not negative impact. Indeed, existing literature on zero-shot translation also suffers from catastrophic forgetting that is common in transformer training (French, 1999; Kirkpatrick et al., 2017).

We therefore rank and select the top few source languages that are closer to the target low resource language using two metrics below.

We rank source languages according to their closeness to the low resource language. We construct the Family of Choice (FAMC) by comparing different ways of ranking linguistic distances empirically based on the small low resource data.

Let $S_s$ and $S_t$ be the source and target sentences, let $L_s$ be the source length, let $P(S_t = s_t | s_s, L_s)$ be the alignment probability, let $F_s$ be the fertility of how many target words a source word is aligned to, let $D_t$ be the distortion based on the fixed distance-based reordering model (Koehn, 2009).

We first construct a word-replacement model based on aligning the small amount of target low resource data with that of each source language using fast_align (Dyer et al., 2013). We replace every source word with the most probable target word according to the product of the alignment probability and the probability of fertility equalling one and distortion equalling zero $P(F_s = 1, D_t = 0 | s_t, s_s, L_s)$. We choose a simple word-replacement model because we aim to work with around 1,000 lines of low resource data. For fast and efficient ranking on such small data, a word-replacement model suits our purpose.

Our distortion measure is the probability of distortion equalling zero, $P(D_t = 0 | s_t, s_s, L_s)$, aggregated over all words in a source language. We use the distortion measure to rank the source languages and obtain the distortion-based FAMC ($FAMD$); we use the translation BLEU scores of the word-replacement model to build the performance-based FAMC ($FAMP$). In Table 1, using Eastern Pokomchi and English as examples, we list the top ten languages in FAMD and FAMP respectively and we star those in FAMO 1.

To prepare for transformer training, we choose the top ten languages neighboring our target low resource language in FAMD and FAMP. We choose ten because existing literature shows that training with two neighboring language families is sufficient to produce quality translation through cross-lingual transfer (Zhou et al., 2018a). Since for some low resource languages, there may not be ten languages in FAMO in our database, we add languages from neighboring families to make an expanded list if such information is known. We use $FAMO^+$ to denote an expanded version of FAMO.

### 3.3 Iterative Pretraining

We have two stages of pretraining using multilingual order-preserving lexiconized transformer on the complete and the star configuration. We design iterative pretraining on symmetric data to address catastrophic forgetting that is common in transformer training (French, 1999; Kirkpatrick et al., 2017).

#### 3.3.1 Stage 1: Pretraining on Neighbors

Firstly, we pretrain on the complete graph configuration of translation paths using the top ten languages neighboring our target low resource language in FAMD, FAMP, and $FAMO^+$ respectively. Low resource data is excluded in training.

We employ the multilingual order-preserving lexiconized transformer for pretraining. Our vocabulary is the combination of the vocabulary for the top ten languages together with the small low resource vocabulary. The final model can translate from any of the ten languages to each other.

#### 3.3.2 Stage 2: Adding Low Resource Data

We include the low resource data in the second stage of training. Since the low resource data covers ~3.5% of the text while all the source languages

---

1In Table 1 and Table 5, Kanjobal is Eastern Kanjobal, Mam is Northern Mam, Cuzco is Cuzco Quechua, Ayacucho is Ayacucho Quechua, Bolivian is South Bolivian Quechua, and Huallaga is Huallaga Quechua.
En terwyl Hy langs die see van Galiléa loop, sien Hy Simon en Andréas, sy broer, besig om 'n net in die see uit te gooi; want hulle was visserne.

And as He drew near to the lake of Galilee, He Simon saw Andrew, and his brother, lying in the lake, for they were fisherman.

En toe Hy daarvandaan 'n bietjie verder gaan, sien Hy Jakobus, die seun van Sebedé¨us, en Johannes, sy broer, wat besig was om die nette in die skuit heel te maak.

And being in a distance, He saw James, the son of Zebedee, and John, his brother, who kept the nets in the boat.

En verder Jakobus, die seun van Sebedé¨us, en Johannes, die broer van Jakobus- aan hulle het Hy die bynaam Boanérges gegee, dit is, seuns van die donder-

And James the son of Zebedee, and John the brother of James; and He gave to them the name, which is called Boun- erges, being of the voice.

| Source Sentence | IPML Translation | Reference |
|-----------------|------------------|-----------|
| En terwyl Hy langs die see van Galiléa loop, sien Hy Simon en Andréas, sy broer, besig om 'n net in die see uit te gooi; want hulle was visserne. | And as He drew near to the lake of Galilee, He Simon saw Andrew, and his brother, lying in the lake, for they were fisherman. | And walking along beside the Sea of Galilee, He saw Simon and his brother Andrew casting a small net in the sea; for they were fishers. |
| En toe Hy daarvandaan 'n bietjie verder gaan, sien Hy Jakobus, die seun van Sebedé¨us, en Johannes, sy broer, wat besig was om die nette in die skuit heel te maak. | And being in a distance, He saw James, the son of Zebedee, and John, his brother, who kept the nets in the boat. | And going forward from there a little, He saw James the son of Zebedee, and his brother John. And they were in the boat mending the nets. |
| En verder Jakobus, die seun van Sebedé¨us, en Johannes, die broer van Jakobus- aan hulle het Hy die bynaam Boanérges gegee, dit is, seuns van die donder- | And James the son of Zebedee, and John the brother of James; and He gave to them the name, which is called Boanerges, being of the voice. | And on James the son of Zebedee, and John the brother of James, He put on them the names Boanerges, which is, Sons of Thunder. |

Table 2: Examples of Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer (IPML) translation from Afrikaans to English as a hypothetical low resource language using FAMP. We train on only 1,093 lines of English data.

4 Data

We use the Bible dataset and the medical EMEA dataset (Mayer and Cysouw, 2014; Tiedemann, 2012). EMEA dataset is from the European Medicines Agency and contains a lot of medical information that may be beneficial to the low resource communities. Our method can be applied to other datasets like WASH guidelines.

For the Bible dataset, we use 124 source languages with 31,103 lines of data and a target low resource language with ∼1,000 lines (∼3.5%) of data. We have two setups for the target low resource language. One uses Eastern Pokomchi, a Mayan language; the other uses English as a hypothetical low resource language. We train on only ∼1,000 lines of low resource data from the book of Luke and test on the 678 lines from the book of Mark. Mark is topically similar to Luke, but is written by a different author. For the first stage of pretraining, we use 80%, 10%, 10% split for training, validation and testing. For the second stage onwards, we use 95%, 5% split of Luke for training and validation, and 100% of Mark for testing.

Eastern Pokomchi is Mayan, and English is Germanic. Since our database does not have ten members of each family, we use FAMO+, the expanded version of FAMO. For English, we include five Germanic languages and five Romance languages in our FAMO+; for Eastern Pokomchi, we include five Mayan languages and five Amerindian languages in our FAMO+. The Amerindian family is largely believed to be close to the Mayan family by the linguistic community.

We construct FAMCs by comparing different ways of ranking linguistic distances empirically.
based on ~1,000 lines of training data. In Table 1, we list the top ten languages for Eastern Pokomchi and English in FAMD and FAMP respectively.

In addition to the Bible dataset, we test on the medical EMEA dataset from the European Medicines Agency (Tiedemann, 2012) to show that our result is useful for other datasets. Using English as a hypothetical language, we train on randomly sampled 1,093 lines of English data, and test on 678 lines of data. Since there are only 9 languages in Germanic and Romance families in EMEA dataset, we include Polish in our FAMO+ for experiments.

There are two key decisions in our data choice. Firstly, we choose to use ~1,000 lines (~3.5%) of low resource data. Most of the multilingual works use equal amount of training data across all languages. Some use low or partial data in low resource languages; however, they require at least between 4K to 100K lines (Lin et al., 2020; Zhou et al., 2018a; Qi et al., 2018) of the low resource data to produce quality translation. Zero-shot translations is promising but also suffer from the small pretraining corpus (Lauscher et al., 2020; Lin et al., 2020; Pfeiffer et al., 2020). Unlike existing works, we aim to improve BLEU and the translation quality to human readable standard using only ~1,000 lines (~3.5%) of low resource data without external dictionary or linguistic helpers.

Secondly, we choose to include Eastern Pokomchi in addition to using English as a hypothetical low resource language. Most research use rich resource languages as hypothetical low resource languages in multilingual experiments (Bird, 2020), few uses real low resource languages (Guzmán et al., 2019). Though data size can be constrained to mimic severely low resource scenarios, much implicit information is still used for the hypothetical low resource language that is actually rich resource. For example, many, including us, use English as a hypothetical low resource language. However, implicit information like English is Germanic is often used. Due to historical colonization, many non-European languages including Southeast Asian languages and African languages are heavily influenced by English. Furthermore, the neighboring languages to the hypothetical low resource language may be rich resource too, rendering more unrealistic treatment in training. Using English as a hypothetical low resource language is therefore not truly low resource. For real low resource scenarios, the family information may have yet to be determined; the neighboring languages may be unknown, and if they are known, they are highly likely to be low resource too. Therefore, we include Eastern Pokomchi.

5 Results
We compare our iteratively pretrained multilingual order-preserving lexiconized transformer (IPML) with five baselines in Table 3. MLc is a baseline model of multilingual order-preserving lexiconized transformer training on complete configuration; in other words, we skip the first stage of pretraining and train on the second stage in Chapter 3.3.2 only. MLs is a baseline model of multilingual order-preserving lexiconized transformer training on star configuration; in other words, we skip both steps of pretraining and train on the final stage in Chapter 3.3.3 only. PMLc is a baseline model of pretrained multilingual order-preserving lexiconized transformer training on complete configuration; in other words, we skip the final stage of training after completing both stages of pretraining. PMLs is a baseline model of pretrained multilingual order-preserving lexiconized transformer training on complete configuration; in other words, we skip both steps of pretraining and train on the final stage in Chapter 3.3.3 only. AML is a baseline model of multilingual order-preserving lexiconized transformer on asymmetric data. We replicate the ~1,000 lines of the low resource data till it matches the training size of other

| Experiments | IPML | MLc | MLs | PMLc | PMLs | AML |
|-------------|------|-----|-----|------|------|-----|
| Pretrained  | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   |
| Iterative   | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   |
| Lexiconized | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   |
| Symmetrical | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   |
| Star        | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   |
| Complete    | ✓    | ✓   | ✓   | ✓    | ✓    | ✓   |

German 35.0 11.6 12.3 33.3 34.5 25.4
Danish 36.0 12.5 12.4 33.3 34.2 26.2
Dutch 35.6 11.5 11.1 32.3 33.7 25.0
Norwegian 35.7 12.3 12.0 33.2 34.1 25.8
Swedish 34.5 11.8 12.4 32.3 33.4 24.9
Spanish 36.4 11.7 11.8 34.1 35.0 26.2
French 35.3 10.8 10.8 33.1 34.0 25.8
Italian 35.9 11.7 11.7 34.3 34.5 26.1
Portuguese 31.5 9.6 10.1 30.0 30.4 23.1
Romanian 34.6 11.3 12.1 32.3 33.2 25.0

Table 3: Comparing our iteratively pretrained multilingual order-preserving lexiconized transformer (IPML) with the baselines training on 1,093 lines of English data in FAMO+. We checkmark the key components used in each experiments and explain all the baselines in details in Section 5.
source languages; we train on the complete graph configuration using eleven languages. Though the number of low resource training lines is the same as others, information is highly asymmetric.

IPML beats the two baselines MLc and MLs that skip pretraining as shown in Table 3. Using English as a hypothetical low resource language training on FAMO+, Spanish-English translation improves from 11.7 (MLc) and 11.8 (MLs) to 36.4 (IPML) with iterative pretraining. Training with the low resource language on both the source and the target sides boosts translation into the target side. Iterative pretraining has a slight advantage over one stage of pretraining. Spanish-English translation improves from 34.1 (PMLc) and 35.0 (PMLs) to 36.4 (IPML). Pretraining is therefore key.

All three pretrained models on symmetric data, IPML, PMLc and PMLs, beat asymmetric baseline AML. In Table 3, we have a +10.2 BLEU increase over our asymmetric baseline on Spanish-English translation using English as a hypothetical low resource language training on FAMO+. All four use the same amount of data, but differ in training strategies and data configuration. In severely low resource scenarios, effective training strategies and symmetric data configuration improves translation.

We compare IPML results training on different sets of source languages in FAMO+, FAMD, and FAMP, for English (our hypothetical low resource language) and Eastern Pokomchi (our real-life low resource language) in Table 4 and 5. Though they consist of different languages, we align common languages for easy comparison. FAMP performs the best for translation from German to English while both FAMP and FAMD outperforms FAMO+ as shown in Table 4. FAMD performs best for translation from Chuj to Eastern Pokomchi as shown in Table 5. When language family information is limited in real-life severely low resource scenarios, constructing FAMC to determine neighbors empirically is very useful in translation.

Comparing Eastern Pokomchi results with English results, we see that translation into real-life severely low-resource languages is more difficult than translation into hypothetical ones. Using English as a hypothetical low resource language, almost all experiments achieve above-30 BLEU scores as shown in Table 4. Using Eastern Pokomchi as our real-life low resource language, most experiments have around-20 BLEU scores as shown in Table 5. Performance for Eastern Pokomchi is not as high as that for English. Indeed, translation into real severely low resource languages is more difficult than hypothetical scenarios.

### Table 4: Performance of Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer (IPML) training on English’s FAMO⁺, FAMD and FAMP. We train on only 1,093 lines of English data.

| Input Language Family | By Linguistics | By Distortion | By Performance |
|-----------------------|----------------|---------------|----------------|
| **FAMO⁺**             | Source BLEU    | Source BLEU   | Source BLEU    |
| German                | 35.0           | German        | 36.7           | German         | 37.6           |
| Danish                | 36.0           | Danish        | 37.1           | Danish         | 37.5           |
| Dutch                 | 35.6           | Dutch         | 35.6           | Dutch          | 36.7           |
| Norwegian             | 35.7           | Norwegian     | 36.9           | Norwegian      | 37.1           |
| Swedish               | 34.5           | Afrikaans     | 38.3           | Afrikaans      | 39.3           |
| Spanish               | 36.4           | Marshallese   | 34.7           | Spanish        | 38.4           |
| French                | 35.3           | French        | 36.0           | French         | 36.6           |
| Italian               | 35.9           | Italian       | 36.9           | Italian        | 37.7           |
| Portuguese            | 31.5           | Portuguese    | 32.9           | Portuguese     | 33.1           |
| Romanian              | 34.6           | Frisian       | 36.1           | Frisian        | 36.9           |

### Table 5: Performance of Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer (IPML) training on Eastern Pokomchi’s FAMO⁺, FAMD and FAMP. We train on only 1,086 lines of Eastern Pokomchi data.

| Input Language Family | By Linguistics | By Distortion | By Performance |
|-----------------------|----------------|---------------|----------------|
| **FAMO⁺**             | Source BLEU    | Source BLEU   | Source BLEU    |
| Chuj                  | 21.8           | Chuj          | 21.9           | Chuj           | 21.6           |
| Cakchiquel            | 22.2           | Cakchiquel    | 22.1           | Cakchiquel     | 21.3           |
| Guajajara             | 19.7           | Guajajara     | 19.1           | Guajajara      | 18.8           |
| Mam                   | 22.2           | Russian       | 22.2           | Mam            | 21.7           |
| Kanjobal              | 21.9           | Toba          | 21.9           | Kanjobal       | 21.4           |
| Cuzco                 | 22.3           | Myanmar       | 19.1           | Thai           | 21.8           |
| Ayacuchu              | 21.6           | Slovensky     | 22.1           | Dadibi         | 19.8           |
| Bolivian              | 22.2           | Latin         | 21.9           | Gumatj         | 19.1           |
| Huallaga              | 22.2           | Ilokano       | 22.5           | Navajo         | 21.3           |
| Aymara                | 21.5           | Norwegian     | 22.6           | Kim            | 21.5           |

### Table 6: Performance of IPML on the EMEA dataset. We train on only 1,086 lines of English data.

| Source | BLEU |
|--------|------|
| German | 34.8 |
| Danish | 37.7 |
| Dutch  | 39.7 |
| Swedish| 37.7 |
| Spanish| 42.8 |
| French | 41.6 |
| Italian| 39.2 |
| Portuguese| 42.8 |
| Romanian| 40.0 |
| Polish | 34.1 |

### Table 7: Testing IPML on the rest of the Bible (30,022 lines). We train on only 1,086 lines of English data.

| Source | BLEU |
|--------|------|
| German | 29.4 |
| Danish | 28.8 |
| Dutch  | 29.9 |
| Norwegian| 29.7 |
| Swedish | 29.0 |
| Spanish | 30.3 |
| French  | 28.9 |
| Italian | 29.7 |
| Portuguese| 24.4 |
| Romanian| 28.8 |
Caso detecte efeitos graves ou outros efeitos não mencionados neste folheto, informe o médico veterinário.

If you notice any side effects or other side effects not mentioned in this leaflet, please inform the veterinarian.

No tratamento de Bovinos com mais de 250 Kg de peso vivo, dividir a dose de forma a não administrar mais de 10 ml por local de injeção.

In the treatment of infants with more than 250 kg in vivo body weight, a the dose to not exceed 10 ml per injection.

For treatment of cattle over 250 kg body weight, divide the dose so that no more than 10 ml are injected at one site.

No entanto, uma vez que é possível a ocorrência de efeitos secundários, qualquer tratamento que exceda as 1-2 semanas deve ser administrado sob supervisão veterinária regular.

However, because any of side effects is possible, any treatment that 1-5 weeks should be administered under regular supervision.

Table 8: Examples of IPML translation on medical EMEA dataset from Portuguese to English using FAMO$^+$.  

| Source Sentence | IPML Translation | Reference |
|-----------------|------------------|-----------|
| Caso detecte efeitos graves ou outros efeitos não mencionados neste folheto, informe o médico veterinário. | If you notice any side effects or other side effects not mentioned in this leaflet, please inform the veterinarian. | If you notice any serious effects or other effects not mentioned in this leaflet, please inform your veterinarian. |
| No tratamento de Bovinos com mais de 250 Kg de peso vivo, dividir a dose de forma a não administrar mais de 10 ml por local de injeção. | In the treatment of infants with more than 250 kg in vivo body weight, a the dose to not exceed 10 ml per injection. | For treatment of cattle over 250 kg body weight, divide the dose so that no more than 10 ml are injected at one site. |

We are curious of how our model trained on ~1,000 lines of data performs on the rest of the Bible. In other words, we would like to know how IPML performs if we train on ~3.5% of the Bible and test on ~96.5% of the Bible. In Table 7, we achieve a BLEU score of 30.3 for Spanish-English translation training IPML on randomly sampled 1,093 lines of data using English as a hypothetical low resource language on FAMO$^+$.  

We show examples of translation into English and Eastern Pokomchi in Table 2 and 9. We observe that the source content is translated well overall and there are a few places for improvement in the translation from Afrikaans to English using 1,093 lines of English data in Table 2. In the first example, “fishermen” and “fishers” are paraphrases of the same concept. IPML successfully predicts the correct concept though it is penalized by BLEU.

Infusing the order-preserving lexiconized component to our training greatly improves qualitative evaluation. But it does not affect BLEU much as BLEU has its limitations in severely low resource scenarios. This means that the BLEU comparison in our paper also applies to the comparison of all experiments without the order-preserving lexiconized component. This is important in real-life situations when a low resource lexicon list is not available, or has to be invented. For example, a person growing up in a local village in Papua New Guinea may have met many people named “Bosai” or “Kaura”, but may have never met a person named “Matthew”, and we may need to create a lexicon word in the low resource language for “Matthew” possibly through phonetics.

We also see good results with the medical EMEA dataset. Treating English as a hypothetical low resource language, we train on only 1,093 lines of English data. For Portuguese-English translation, we obtain a BLEU score of 42.8 while the rest of languages all obtain BLEU scores above 34 in Table 6 and Table 8. This shows that our method is useful for other datasets.

6 Conclusion

We translate a closed text that is known in advance into a severely low resource language by leveraging massive source parallelism. We present two empirical metrics to rank the 124 source languages to determine their closeness to the low resource language and construct FAMCs. We build an iteratively pretrained multilingual order-preserving lexiconized transformer to train on ~1,000 lines of low resource data. We obtain a +24.7 BLEU increase over a multilingual transformer baseline (MLc) in Spanish-English translation using English as a hypothetical low resource language in the Bible dataset. By exploring data symmetry, we have a +10.2 BLEU increase over our asymmetric baseline. Our results for a real severely low resource language, Eastern Pokomchi, show improvements. Finally, we add a multilingual order-preserving lexiconized component to translate the named entities accurately. We build a massively parallel lexicon table for 2,939 Bible named entities in 124 source languages, covering more than 66 severely low resource languages. Our result for the medical EMEA dataset shows that our method has potential use for a broad universe of datasets and applications.

Our final result can also serve as a ranking metric for linguistic distances though it is much more expensive in terms of time and resources. In the future, we would like to explore more metrics that are fast and efficient in ranking linguistic distances to the severely low resource language.
References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 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.

Ife Adebara, El Moatez Billah Nagoudi, and Muhammad Abdul Mageed. 2020. Translating similar languages: Role of mutual intelligibility in multilingual transformers. arXiv preprint arXiv:2011.05037.

Antonios Anastasopoulos, Sameer Bansal, David Chiang, Sharon Goldwater, and Adam Lopez. 2017. Spoken term discovery for language documentation using translations. In Proceedings of the Workshop on Speech-Centric Natural Language Processing, pages 53–58.

Sameer Bansal, Herman Kamper, Sharon Goldwater, and Adam Lopez. 2017. Weakly supervised spoken term discovery using cross-lingual side information. In Acoustics, Speech and Signal Processing, pages 5760–5764. IEEE.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. arXiv preprint arXiv:2004.14928.

Alexandre Bérard, Zae Myung Kim, Vassilina Nikouлина, Eunjeong Lucy Park, and Matthias Gallé. 2020. A multilingual neural machine translation model for biomedical data. In Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 745–754.

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

John S Brownstein, Clark C Freifeld, Ben Y Reis, and Kenneth D Mandl. 2008. Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. PLoS medicine, 5(7):e151.

Jasone Cenoz. 2001. The effect of linguistic distance, L2 status and age on cross-linguistic influence in third language acquisition. Cross-linguistic influence in 2nd language acquisition: Psycholinguistic perspectives, 111(45):8–20.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2020. Reusing a pretrained language model on languages with limited corpora for unsupervised nmt. arXiv preprint arXiv:2009.07610.

Bernard Comrie. 2005. The world atlas of language structures. Oxford University Press.

Boele De Raad, Marco Perugini, and Zsófia Szirmák. 1997. In pursuit of a cross-lingual reference structure of personality traits: Comparisons among five languages. European Journal of Personality, 11(3):167–185.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. arXiv preprint arXiv:2004.03672.

Haim Dubossarsky, Ivan Vulic, Roi Reichart, and Anna korhonen. 2020. The secret is in the spectra: Predicting cross-lingual task performance with spectral similarity measures. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2377–2390.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In Proceedings of the 12th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, pages 644–648.

Paul S Earle, Daniel C Bowden, and Michelle Guy. 2012. Twitter earthquake detection: earthquake monitoring in a social world. Annals of Geophysics, 54(6).

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 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.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. Cognition, 28(1):3–71.

Markus Freitag and Orhan Firat. 2020. Complete multilingual neural machine translation. arXiv preprint arXiv:2010.10239.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. Trends in cognitive sciences, 3(4):128–135.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 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.

Raymond G Gordon Jr. 2005. Ethnologue, languages of the world. http://www.ethnologue.com/.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. arXiv preprint arXiv:1410.5401.
7 Appendix

We show an illustration of WASH guidelines in Figure 2. We also show IPML translations into Eastern Pokomchi (Mayan) in Table 9.
Figure 2: An Amharic illustration of translation of water, sanitation, and hygiene (WASH) guidelines in Ethiopia (USAID, 2009).

| Source Sentence | IPML Translation | Reference |
|-----------------|------------------|-----------|
| Ket idi limmabas iti dinna ti baybay ti Galilea, nakitana ni Simon ken ni Andres a cabsatna, nga iwaywayatda ti iket iti baybay; ta dumadaclisida idi. | Eh noq ojik i rub'an i Jesús juntar i k'isa palaw i Galilea, xrilow reje i Simón ruch'ihil i Andres, re' i rutuut i k'isa palaw, ruum jinaj i k'isa palaw barco. | Noq k'ahch'i rik'ik i Jesús chi chi'i i k'isa palaw ar Galilea, xrilow reje wach i Simón ruch'ihil i ruchhaaq'. Andres rub'ihnaal. Re' keh aj karineel taqe, k'ahch'i kikutum qohoq i kiya'l pan palaw. |
| Ket idi nagna pay bassit nakitana ni Santiago nga anac ni Zebedeo ken ni Juan a cabsatna, nga addada idi iti barangayda, a tartarimaanenda dag-iti iketda. | Eh noq ojik i rub'an i Jesús, xrilow i Jacobo, re' i Jacobo rak'uun i Zebedeo, re' Juan rub'ihnaal, ruch'ihil taqe i raj tahqaneel. eh xkikoj wo' wach chinh'a i k'isa palaw. | Eh junk'aam-oq chik i xb'ehik reje i Jesús, xrilow kiwach i k'ib' chi winaq kicha'aq' k'ib', re' Jacobo, re' Juan, rak'uun taqe i Zebedeo. Eh wilkeeb' chupaam jinaj i barco, k'ahch'i kik'ojem wach i kiya'l b'amb'al kar. |
| Ket immasideg ni Jesus ket inigga- manna iti imana ket pinatacderna; ket pinanawan ti gorigor , ket nagservi cadacuada. | Eh re' Jesús xujil i koq riib', xutz'aj' i koq chinaah i q'ab'. eh re' i kaq tz'a' chi rii. eh jumehq'uil xwuktiik joohtoq, re' chik i reh xutoq'aa' cho yej-anik kiwa'. | Eh re' i Jesús xujil i koq riib' ruuk' i yowaab', xuchop chi q'ab', xruksaj joh-toq, eh jumehq'uil xik'ik i tz'a' chi rii. Eh re' chik i reh xutoq'aa' cho yej-anik kiwa'. |

Table 9: Examples of Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer (IPML) translation from Ilokano to Eastern Pokomchi using FAMD. We train on only 1,086 lines of Eastern Pokomchi data.