Massively Parallel Cross-Lingual Learning in Low-Resource Target Language Translation

Zhong Zhou
Carnegie Mellon University
zhongzhou@cmu.edu

Matthias Sperber
Karlsruhe Institute of Technology
matthias.sperber@kit.edu

Alex Waibel
Carnegie Mellon University
Karlsruhe Institute of Technology
alex@waibel.com

Abstract

We work on translation from rich-resource languages to low-resource languages. The main challenges we identify are the lack of low-resource language data, effective methods for cross-lingual transfer, and the variable-binding problem that is common in neural systems. We build a translation system that addresses these challenges using eight European language families as our test ground. Firstly, we add the source and the target family labels and study intra-family and inter-family influences for effective cross-lingual transfer. We achieve an improvement of +9.9 in BLEU score for English-Swedish translation using eight families compared to the single-family multi-source multi-target baseline. Moreover, we find that training on two neighboring families closest to the low-resource language is often enough. Secondly, we construct an ablation study and find that reasonably good results can be achieved even with considerably less target data. Thirdly, we address the variable-binding problem by building an order-preserving named entity translation model. We obtain 60.6% accuracy in qualitative evaluation where our translations are akin to human translations in a preliminary study.

1 Introduction

We work on translation from a rich-resource language to a low-resource language. There is usually little low-resource language data, much less parallel data available (Duong et al., 2016; Anastasopoulos et al., 2017); Despite of the challenges of little data and few human experts, it has many useful applications. Applications include translating water, sanitation and hygiene (WASH) guidelines to protect Indian tribal children against waterborne diseases, introducing earthquake preparedness techniques to Indonesian tribal groups living near volcanoes and delivering information to the disabled or the elderly in low-resource language communities (Reddy et al., 2017; Barrett, 2005; Anastasiou and Schäler, 2010; Perry and Bird, 2017). These are useful examples of translating a closed text known in advance to the low-resource language.

There are three main challenges. Firstly, most of previous works research on individual languages instead of collective families. Cross-lingual impacts and similarities are very helpful when there is little data in low-resource language (Shoemark et al., 2016; Sapir, 1921; Odlin, 1989; Cenoz, 2001; Toral and Way, 2018; De Raad et al., 1997; Hermans, 2003; Specia et al., 2016). Secondly, most of the multilingual Neural Machine Translation (NMT) works assume the same amount of training data for all languages. In the low-resource case, it is important to exploit low or partial data in low-resource language to produce high quality translation. The third issue is the variable-binding problem that is common in neural systems, where "John calls Mary" is treated the same way as "Mary calls John" (Fodor and Pylyshyn, 1988; Graves et al., 2014). It is more challenging when both "Mary" and "John" are rare words. Solving the binding problem is crucial because the mistakes in named entities change the meaning of the translation. It is especially challenging in the low-resource case because many words are rare words.

Our contribution in addressing these issues is three-fold, extending from multi-source multi-target attentional NMT. Firstly, to examine intra-family and inter-family influences, we add source and target language family labels in training. Training on multiple families improves BLEU score significantly; moreover, we find training on two neighboring families closest to the low-resource language gives reasonably good BLEU scores, and we define neighboring families closely in Section 3.2. Secondly, we conduct an ablation

Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 232–243

https://doi.org/10.18653/v1/W18-6324

study to explore how generalization changes with different amounts of data and find that we only need a small amount of low-resource language data to produce reasonably good BLEU scores. We use full data except for the ablation study. Finally, to address the variable-binding problem, we build a parallel lexicon table across twenty-three European languages and devise a novel method of order-preserving named entity translation method. Our method works in translation of any text with a fixed set of named entities known in advance. Our goal is to minimize manual labor, but not to fully automate to ensure the correct translation of named entities and their ordering.

In this paper, we begin with introduction and related work in Section 1 and 2. We introduce our methods in addressing three issues that are important for translation into low-resource language in Section 3.2, as proposed extensions to our baseline in Section 3.1. Finally, we present our results in Section 4 and conclude in Section 5.

2 Related Work

2.1 Multilingual Attentional NMT

Attentional NMT is trained directly in an end-to-end system and has flourished recently (Wu et al., 2016; Sennrich et al., 2016; Ling et al., 2015). Machine polyglotism, training machines to be proficient in many languages, is a new paradigm of multilingual NMT (Johnson et al., 2017; Ha et al., 2016; Firat et al., 2016; Zoph and Knight, 2016; Dong et al., 2015; Gillick et al., 2016; Al-Rfou et al., 2013; Tsvetkov et al., 2016). Many multilingual NMT systems involve multiple encoders and decoders, and it is hard to combine attention for quadratic language pairs bypassing quadratic attention mechanisms (Firat et al., 2016). In multi-source scenarios, multiple encoders share a combined attention mechanism (Zoph and Knight, 2016). In multi-target scenarios, every decoder handles its own attention with parameter sharing (Dong et al., 2015). Attention combination schemes include simple combination and hierarchical combination (Libovickỳ and Helcl, 2017).

The state-of-the-art of multilingual NMT is adding source and target language labels in training a universal model with a single attention scheme, and Byte-Pair Encoding (BPE) is used at preprocessing stage (Ha et al., 2016). This method is elegant in its simplicity and its advancement in low-resource language translation as well as zero-shot translation using pivot-based translation scheme (Johnson et al., 2017). However, these works have training data that increases quadratically with the number of languages (Dong et al., 2015; Gillick et al., 2016), rendering training on massively parallel corpora difficult.

2.2 Sub-word Level NMT

Many NMT systems lack robustness with out-of-vocabulary words (*OOV*s) (Wu et al., 2016). Most *OOV*s are treated as unknowns (*$UNK*s) uniformly, even though they are semantically important and different (Ling et al., 2015; Sennrich et al., 2016). To tackle the *OOV* problem, researchers work on byte-level (Gillick et al., 2016) and character-level models (Ling et al., 2015; Chung et al., 2016). Many character-level models do not work as well as word-level models, and do not produce optimal alignments (Tiedemann, 2012). As a result, many researchers shift to subword level modeling between character-level and word-level. One prominent direction is BPE which iteratively learns subword units and balances sequence length and expressiveness with robustness (Sennrich et al., 2016).

2.3 Lexiconized NMT

Much research is done in translating lexicons and named entities in NMT (Nguyen and Chiang, 2017; Wang et al., 2017; Arthur et al., 2016). Some researchers create a separate character-level named entity model and mark all named entities as *$TERM*s to train (Wang et al., 2017). This method learns people's names well but does not improve BLEU scores (Wang et al., 2017). It is time-consuming and adds to the system complexity. Other researchers attempt to build lexicon translation seamlessly with attentional NMT by using an affine transformation of attentional weights (Nguyen and Chiang, 2017; Arthur et al., 2016). Some also attempt to embed cross-lingual lexicons into the same vector space for transfer of information (Duong et al., 2017).

3 Translation System

3.1 Baseline Translation System

Our baseline is multi-source multi-target attentional NMT within one language family through adding source and target language labels with a single unified attentional scheme, with BPE used at the preprocessing stage. The source and target vocabulary are not shared.

Families	Languages
Germanic	German (de) Danish (dn) Dutch (dt) Norwegian (no) Swedish (sw) English (en)
Slavic	Croatian (cr) Czech (cz) Polish (po) Russian (ru) Ukrainian (ur) Bulgarian (bg)
Romance	Spanish (es) French (fr) Italian (it) Portuguese (po) Romanian (ro)
Albanian	Albanian (ab)
Hellenic	Greek (gk)
Italic	Latin (ln) [descendants: Romance languages]
Uralic	Finnish (fn) Hungarian (hg)
Celtic	Welsh (ws)

Table 1: Language families. Language codes are in brackets.

3.2 Proposed Extensions

We present our methods in solving three issues relevant to translation into low-resource language as our proposed extensions.

3.2.1 Language Families and Cross-lingual Learning

Cross-lingual and cross-cultural influences and similarities are important in linguistics (Shoemark et al., 2016; Levin et al., 1998; Sapir, 1921; Odlin, 1989; Cenoz, 2001; Toral and Way, 2018; De Raad et al., 1997; Hermans, 2003; Specia et al., 2016). The English word, "Beleaguer" originates from the Dutch word "belegeren"; "fidget" originates from the Nordic word "fikja". English and Dutch belong to the same family and their proximity has effect on each other (Harding and Sokal, 1988; Ross et al., 2006). Furthermore, languages that do not belong to the same family affect each other too (Sapir, 1921; Ammon, 2001; Toral and Way, 2018). "Somatic" stems from the Greek word "soma"; "広告" (Japanese), "광고"(Korean), "Quảng cáo"(Vietnamese) are closely related to the Traditional Chinese word "廣告". Indeed, many cross-lingual similarities are present.

In this paper, we use the language phylogenetic tree as the measure of closeness of languages and language families (Petroni and Serva, 2008). The distance measure of language families is the collective of all of the component languages. Language families that are next to each other in the language phylogenetic tree are treated as neighboring families in our paper, like Germanic family and Romance family. In our discussion in this paper, we will often refer to closely related families in the language phylogenetic tree as neighboring families.

We prepend the source and target family labels, in addition to the source and target language labels to the source sentence to improve convergence rate and increase translation performance. For example, all French-to-English translation pairs are prepended with four labels, the source and target family labels and the source and target languages labels, i.e., `__opt_family_src_romance __opt_family_tgt_germanic __opt_src_fr __opt_tgt_en`. In Section 4, we examine intra-family and inter-family effects more closely.

3.2.2 Ablation Study on Target Training data

To achieve high information transfer from rich-resource language to low-resource target language, we would like to find out how much target training data is needed to produce reasonably good performance. We vary the amount of low-resource training data to examine how to achieve reasonably good BLEU score using limited low-resource data. In the era of Statistical Machine Translation (SMT), researchers have worked on data sampling and sorting measures (Eck et al., 2005; Axelrod et al., 2011).

To rigorously determine how much low-resource target language is needed for reasonably good results, we do a range of control experiments by drawing samples from the low-resource language data randomly with replacement and duplicate them if necessary to ensure all experiments carry the same number of training sentences. We keep the amount of training data in rich-resource languages the same, and vary the amount of training data in low-resource language to conduct rigorous control experiments. Our data selection process is different from prior research in that only the low-resource training data is reduced, simulating the real world scenario of having little data in low-resource language. By comparing results from control experiments, we determine how much low-resource data is needed.

3.2.3 Order-preserving Lexiconized NMT

The variable-binding problem is an inherent issue in connectionist architectures (Fodor and Pylyshyn, 1988; Graves et al., 2014). "John calls Mary" is not equivalent to "Mary calls John", but neural networks cannot distinguish the two easily (Fodor and Pylyshyn, 1988; Graves et al., 2014). The failure of traditional NMT to distinguish the subject and the object of a sentence is detrimental. For example, in the narration "John told his son Ryan to help David, the brother of Mary", it is a serious mistake if we reverse John and Ryan's father-son relationships or confuse Ryan's and David's

lan	de	dn	dt	en	no	sw
de	N.A.	37.5	43.4	45.1	41.1	35.8
dn	39.0	N.A.	37.1	41.1	42.6	37.4
dt	43.5	36.3	N.A.	45.1	39.0	34.3
en	40.4	34.5	41.1	N.A.	37.1	34.0
no	40.5	42.7	40.4	42.8	N.A.	40.6
sw	39.4	38.9	37.5	40.4	43.0	N.A.

Table 2: (Baseline model) Germanic family multi-source multi-target translation. Each row represents source, each column represents target. Language codes follow Table 1.

relationships with Mary.

In our research on translation, we focus mainly on text with a fixed set of named entities known in advance. We assume that experts help to translate a given list of named entities into low-resource language first before attempting to translate any text. Under this assumption, we propose an order-preserving named entity translation mechanism. Our solution is to first create a parallel lexicon table for all twenty-three European languages using a seed English lexicon table and fast-aligning it with the rest (Dyer et al., 2013). Instead of using *$UNK*s to replace the named entities, we use *$NE*s to distinguish them from the other unknowns. We also sequentially tag named entities in a sentence as *$NE1*, *$NE2*, . . . , to preserve their ordering. For every sentence pair in the multilingual training, we build a target named entity decoding dictionary by using all target lexicons from our lexicon table that matches with those appeared in the source sentence. During the evaluation stage, we replace all the numbered *$NE*s using the target named entity decoding dictionary to present our final translation. This method improves translation accuracy greatly and preserves the order.

As a result of our contribution, the experts only need to translate a few lexicons and a small amount of low-resource text before passing the task to our system to obtain good results. Post-editing and minor changes may be required to achieve 100% accuracy before the releasing the translation to the low-resource language communities.

4 Experiments and Results

We choose the Bible corpus as a test ground for our proposed extensions because the Bible is the most translated text that exists and is freely accessible. Though it has limitations, it does not have copyright issues like most of literary works that are translated into many languages do. There are many research works done using the Bible (Naaijer and Roorda, 1993; Mayer and Cysouw, 2014; Scannell, 2006; Dufter and Schütze, 2018; Resnik et al., 1999; Chan and Pollard, 2001; Banchs and Costa-Jussà, 2011; Christodouloupoulos and Steedman, 2015; Beale et al., 2005). Unlike many past research works where only New Testament is used (Dufter and Schütze, 2018), we use both Old Testament and New Testament in our Bible corpus. We align all Bible verses across different languages.

We train our proposed model on twenty-three European languages across eight families on a parallel Bible corpus. For our purpose, we treat Swedish as our hypothetical low-resource target language, English as our rich-resource language in the single-source single-target case and all other Germanic languages as our rich-resource languages in the multi-source multi-target case.

Firstly, we present our data and training parameters. Secondly, we add family tags in different configurations of language families showing intra-family and inter-family effects. Thirdly, we conduct an ablation study and plot the generalization curves by varying the amount of training data in Swedish, and we show that training on one fifth of the data give reasonably good BLEU scores. Lastly, we devise an order-preserving lexicon translation method by building a parallel lexicon table across twenty-three European languages and tagging named entities in order.

4.1 Data and Training Parameters

We clean and align the Bible in twenty-three European languages in Table 1. We randomly sample the training, validation and test sets according to the 0.75, 0.15, 0.10 ratio. Our training set contains 23K verses, but is massively parallel. In our control experiments, we also use the experiment training on the WMT'14 French-English dataset together with French and English Bibles to compare with our results. Note that our WMT baseline contains French and English Bibles in addition to the WMT'14 data, and is used to contrast our results with the effect of increasing data.

In all our experiments, we use a minibatch size of 64, dropout rate of 0.3, 4 RNN layers of size 1000, a word vector size of 600, learning rate of 0.8 across all LSTM-based multilingual experiments. For single-source single-target translation, we use 2 RNN layers of size 500, a word vector size of 500, and learning rate of 1.0. All learning rates are decaying at the rate of 0.7 if the validation score is not improving or it is past epoch 9. We use SGD as our learning algorithm. We build our

expt	S	G	GS	GR	3F	8F
de2sw	4.0	35.8	42.0	42.2	42.5	42.8
dn2sw	16.9	37.4	43.4	41.8	42.7	41.7
dt2sw	4.8	34.3	41.4	41.6	42.8	42.5
en2sw	6.9	34.0	40.3	40.2	41.8	42.1
no2sw	16.8	40.6	43.6	44.0	44.5	43.1

Table 3: Inter-family and intra-family effects on BLEU scores with respect to increasing addition of language families.
S: single-source single-target NMT.
G: training on Germanic family.
GS: training on Germanic, Slavic family.
GR: training on Germanic, Romance family.
3F: training on Germanic, Slavic, Romance family.
8F: training on all 8 European families together.

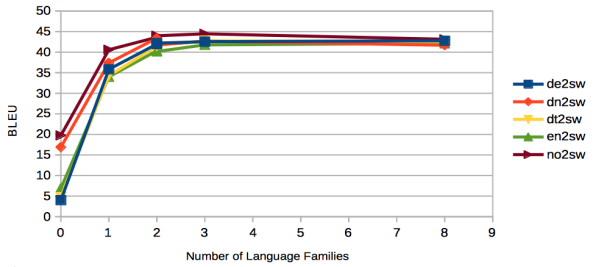

Figure 1: Intra-family and inter-family effects on BLEU scores with respect to increasing addition of language families.

code based on OpenNMT (Klein et al., 2017). For the ablation study, we train on BLEU scores directly until the *Generalization Loss* (*GL*) exceeds a threshold of $\alpha = 0.1$ (Prechelt, 1998). *GL* at epoch t is defined as $GL(t) = 100(1 - \frac{E^t_{val}}{E^t_{opt}})$, modified by us to suit our objective using BLEU scores (Prechelt, 1998). E^t_{val} is the validation score at epoch t and E^t_{opt} is the optimal score up to epoch t. We evaluate our models using both BLEU scores (Papineni et al., 2002) and qualitative evaluation.

4.2 Family labels and Intra-family & Inter-family Effects

We first investigate intra-family and inter-family influences and the effects of adding family labels. We use full training data in this subsection. Adding family labels not only improves convergence rate, but also increases BLEU scores.

Languages have varying closeness to each other: Single-source single-target translations of different languages in Germanic family to Swedish show huge differences in BLEU scores as shown in Table 3. These differences are well aligned with the multi-source multi-target results. Norwegian-Swedish and Danish-Swedish translations have much higher BLEU scores than the rest. This hints that Norwegian and Danish are closer to Swedish than the rest in the neural representation.

Multi-source multi-target translation improves greatly from single-source single-target translation: English-Swedish single-source single-target translation gives a low BLEU score of 6.9 as shown in Table 3, which is understandable as our dataset is very small. BLEU score for English-Swedish translation improves greatly to 34.0 in multi-source multi-target NMT training on Germanic family as shown in Table 2. In this paper, we treat Germanic multi-source multi-target NMT as our baseline model. Complete tables of multi-source and multi-target experiments are in the appendices. We present only relevant columns important for cross-lingual learning and translation into low-resource language here.

expt	S	G	GSl	GRl	3Fl	8Fl
de2sw	4.0	35.8	41.8	42.2	42.5	44.3
dn2sw	16.9	37.4	43.0	41.5	42.5	42.8
dt2sw	4.8	34.3	41.4	41.8	42.7	42.3
en2sw	6.9	34.0	40.9	40.4	41.7	43.9
no2sw	16.8	40.6	43.7	44.3	44.2	44.7

Table 4: Effects of adding family labels on BLEU scores with respect to increasing addition of language families.
S and G: same as in Table 3.
GSl: Germanic, Slavic family with family labels.
GRl: Germanic, Romance family with family labels.
3Fl: Germanic, Slavic, Romance family with family labels.
8Fl: all 8 European families together with family labels

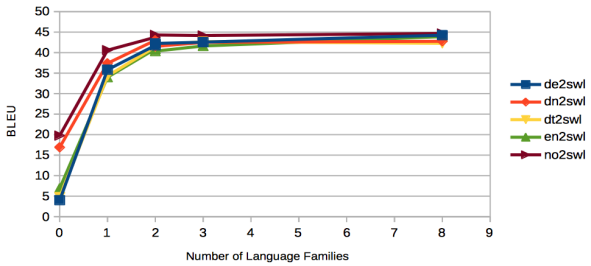

Figure 2: Effects of adding family labels on BLEU scores with respect to increasing addition of language families.

Adding languages from other families into training improves translation quality within each family greatly: English-Swedish translation's BLEU score improves significantly from 34.0 to 40.3 training on Germanic and Slavic families, and 40.2 training on Germanic and Romance families as shown in Table 3. After we add all three families in training, BLEU score for English-Swedish translation increases further to 41.8 in Table 3. Finally, after we add all eight families, BLEU score for English-Swedish translation increases to 42.1 in Table 3.

A Plateau is observed after adding more than one neighboring family: A plateau is observed when we plot Table 3 in Figure 1. The increase in BLEU scores after adding two families is much milder than that of the first addition of a neighbor-

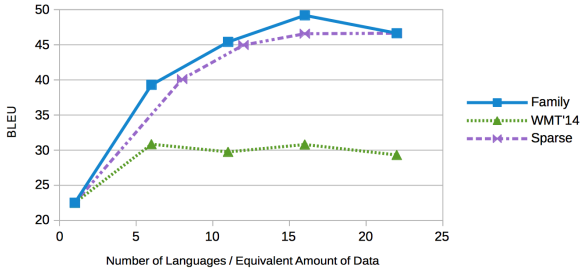

Figure 3: Comparison of different ways of increasing training Data in French-English translation.
Family: Adding data from other languages based on the family unit
WMT'14: Adding WMT'14 data as control experiment
Sparse: Adding data from other languages that spans the eight European families

ing family. This hints that using unlimited number of languages to train may not be necessary.

Adding family labels not only improves convergence rate, but also increases BLEU scores: We observe in Table 4 that BLEU scores for most language pairs improve with the addition of family labels. Training on eight language families, we achieve a BLEU score of 43.9 for English-Swedish translation, +9.9 above the Germanic baseline. Indeed, the more families we have, the more helpful it is to distinguish them.

Training on two neighboring families nearest to the low-resource language gives better result than training on languages that are further apart: Our observation of the plateau hints that training on two neighboring families nearest to the low-resource language is good enough as shown in Table 3. Before jumping to conclusion, we compare results of adding languages by family with that of adding languages by random samples that span all eight families, defined as the following.

Definition 4.1 (Language Spanning)**.** A set of languages spans a set of families when it contains at least one language from each family.

In Figure 3, we conduct a few experiments on French-English translation using different ways of adding training data. Let *family addition* describe the addition of training data through adding closeby language families based on the unit of family; let *sparse addition* describe the addition of training data through adding language sets that spans eight language families. In sparse addition, languages are further apart as each may represent a different family. We find that family addition gives better generalization than that of sparse addition. It strengthens our earlier results that training on two families closest to our low-resource language is a reliable way to reach good generalization.

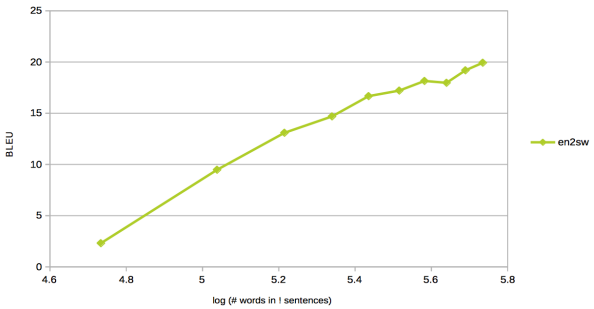

Figure 4: Single-source single-target English-Swedish BLEU plots against increasing amount of Swedish data.

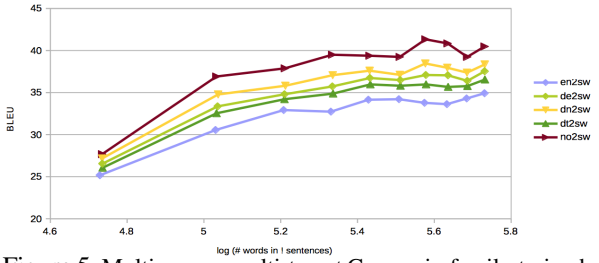

Figure 5: Multi-source multi-target Germanic-family-trained BLEU plots against increasing amount of Swedish data.

Generalization is not merely an effect of increasing amount of data: In Figure 3, we compare all methods of adding languages against a WMT'14 curve by using equivalent amount of WMT'14 French-English data in each experiment. The WMT'14 curve serve as our benchmark of observing the effect of increasing data, we observe that our addition of other languages improve BLEU score much sharply than the increase in the benchmark, showing that our generalization is not merely an effect of increasing data. We also observe that though increase WMT'14 data initially increases BLEU score, it reaches a plateau and adding more WMT'14 data does not increase performance from very early point.

4.3 Ablation Study on Target Training Data

We use full training data from all rich-resource languages, and we vary the amount of training data in Swedish, our low-resource language, spanning from one tenth to full length uniformly. We duplicate the subset to ensure all training sets, though having a different number of unique sentences, have the same number of total sentences.

Power-law relationship is observed between the performance and the amount of training data in low-resource language: Figure 5 shows how BLEU scores vary logarithmically with the number of unique sentences in the low-resource training data. It follows a linear pattern for single-source single-target translation from English to

Data	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
#w	53589	107262	161332	214185	268228	322116	375439	429470	483440	538030
log#w	4.73	5.03	5.21	5.33	5.43	5.51	5.57	5.63	5.68	5.73
en2sw	25.2	30.6	32.9	32.7	34.2	34.2	33.8	33.6	34.3	34.9
de2sw	26.5	33.4	34.8	35.7	36.7	36.5	37.1	37.1	36.4	37.5
dn2sw	27.2	34.8	35.8	37.1	37.6	37.1	38.5	38.0	37.4	38.4
dt2sw	26.1	32.5	34.2	34.9	36.0	35.8	36.0	35.7	35.8	36.6
no2sw	27.7	36.9	37.9	39.5	39.4	39.2	41.3	40.8	39.2	40.5

Table 5: Ablation Study on Germanic Family. #w is the word count of unique sentences in Swedish data.

en	de	cz	es	fn	sw
Joseph	Joseph	Jozef	José	Joseph	Josef
Peter	Petrus	Petr	Pedro	Pietari	Petrus
Zion	Zion	Sion	Sion	Zionin	Sion
John	Johannes	Jan	Juan	Johannes	Johannes
Egypt	Ägypten	Egyptské	Egipto	Egyptin	Egyptens
Noah	Noah	Noé	Noé	Noa	Noa

Table 6: A few examples from the parallel lexicon table.

expt	G	OG	OG1	OGM
de2sw	35.8	36.6	36.6	36.9
dn2sw	37.4	37.0	37.2	36.9
dt2sw	34.3	35.8	35.6	35.9
en2sw	34.0	33.6	33.9	33.4
no2sw	40.6	41.2	41.0	41.4

Table 7: Summary of order-preserving lexicon translation.
G: training on Germanic family without using order-preserving method.
OG: order-preserving lexicon translation.
OG1: OG translation using lexicons with frequency 1.
OGM: OG translation using lexicons with manual selection.

Swedish as shown in Figure 4. We also observe a linear pattern for the multi-source multi-target case, though more uneven in Figure 5. The linear pattern with BLEU scores against the logarithmic data shows the power-law relationship between the performance in translation and the amount of low-resource training data. Similar power-law relationships are also found in past research and contemporary literature (Turchi et al., 2008; Hestness et al., 2017).

We achieve reasonably good BLEU scores using one fifth of random samples: For the multi-source multi-target case, we find that using one fifth of the low-resource training data gives reasonably good BLEU scores as shown in Figure 5. This is helpful when we have little low-resource data. For translation into low-resource language, the experts only need to translate a small amount of seed data before passing it to our system [1].

4.4 Order-preserving Lexiconized NMT

We devise a mechanism to build a parallel lexicon table across twenty-three European languages using very little data and zero manual work. A few lexicon examples are shown in Table 6. We first extract named entities from the English Bible (Manning et al., 2014) and combine them with English biblically named entities from multiple sources (Easton, 1897; Nave, 1903; Smith et al., 1967; Hitchcock, 1874; Rice, 2015). Secondly, we carefully automate the filtering process to obtain a clean English lexicon list. Using this list as the seed, we build a parallel lexicon table across all twenty-three languages through fast-aligning (Dyer et al., 2013). The final parallel lexicon table has 2916 named entities. In the translation task into low-resource language, we assume that the experts first translate these lexicon entries, and then translate approximately one fifth random sentences before we train our NMT. If necessary, the experts evaluate and correct translations before releasing the final translations to the low-resource language community. We aim to reduce human effort in post-editing and increase machine accuracy. After labeling named entities in each sentence pair in order, we train and obtain good translation results.

We observe 60.6% accuracy in human evaluation where our translations are parallel to human translations: In Table 8, we show some examples of machine translated text, we also show the expected correct translations for comparison. Not only the named entities are correctly mapped, but also the ordering of the subject and the object is preserved. In a subset of our test set, we conduct human evaluation on 320 English-Swedish results to rate the translations into three categories: accurate (parallel to human translation), almost accurate (needing minor corrections) and inaccurate. More precisely, each sentence is evaluated using three criteria: correct set of named entities, correct positioning of named entities, and accurate meaning of overall translation. If a sentence achieves all three, then it is termed as accurate; if either a name entity is missing or its position is wrong, then it is termed as almost accurate (needing minor cor-

[1] Note that using nine tenth of random samples yields higher performance than using full data, but it may not be generalized to other datasets.

Source Sentence	NMT Translation without Order Preservation (Before)	NMT Translation with Order Preservation (After)	Correct Target Translation	Frequency of Named Entities
And *Noah* fathered three sons, *Shem*, *Ham*, and *Japheth*.	Och *Noa* födde tre söner, *Sem*, *Ham* och *Jafet*.	Och *Noa* födde tre söner, *Sem Ham* och *Jafet*	Och *Noa* födde tre söner: *Sem*, *Ham* och *Jafet*.	*Noah*: 58, *Shem*: 18, *Ham*: 17, *Japheth*: 11
And *Saul* spoke to his son *Jonathan*, and to all his servants, to kill *David*.	Och *Saul* sade till *Jonatan*, hans son, och alla hans tjänare, så att de skulle döda *David*.	Och *Saul* talade till sin son *Jonatan* och alla hans tjänare för att döda *David*	Och *Saul* talade med sin son *Jonatan* och med alla sina tjänare om att döda *David*	*Saul*: 424, *Jonathan*: 121, *David*: 1134
And they killed *Parshandatha*, and *Dalphon*, and *Aspatha*, and *Poratha*, and *Adalia*, and *Aridatha*, and *Parmashta*, and *Arisai*, and *Aridai*, and *Vajezatha*,	Och de dräpte *Kedak*, *Ir-Fittim*, *Aquila*, *dörrvaktarna*, *Amarja*, *Bered*, *vidare Bet-Hadt*, *Berota*, *Gat-Rimmon*,	Och de dräpte *Parsandata Dalefon* och *Aspata Porata Adalja Aridata Parmasta Arisai Aridai Vajsata*	Och *Parsandata*, *Dalefon*, *Aspata*, *Porata*, *Adalja*, *Aridata*, *Parmasta*, *Arisai*, *Aridai* och *Vajsata*,	*Parshandatha*: 1, *Dalphon*: 1, *Aspatha*: 1, *Poratha*: 1, *Adalia*: 1, *Aridatha*: 1, *Parmashta*: 1, *Arisai*: 1, *Aridai*: 1, *Vajezatha*: 1

Table 8: Examples of order-preserving lexicon-aware translation for English to Swedish. The frequency of the named entities are the number of occurrences each named entity appears in the whole dataset; for example, all named entities in the last sentence only appear in the test set once, and do not appear in the training data.

rection); if the meaning of the sentence is entirely wrong, then it is inaccurate. Our results are 60.6% accurate, 33.8% needing minor corrections, and 5.6% inaccurate. Though human evaluation carries bias and the sample is small, it does give us perspective on the performance of our model.

Order-preservation performs well especially when the named entities are rare words: In Table 8, NMT without order-preservation lexiconized treatment performs well when named entities are common words, but fails to predict the correct set of named entities and their ordering when named entities are rare words. The last column shows the number of occurrences of each named entity. For the last example, there are many named entities that only occur in data once, which means that they never appear in training and only appear in the test set. The normal NMT without order-preservation lexiconized treatment predicts the wrong set of named entities with the wrong ordering. Our lexiconized order-preserving NMT, on the contrary, performs well at both the head and tail of the distribution, predicts the right set of named entities with the right ordering.

Prediction with longer sentences and many named entities are handled well: In Table 8, we see that normal NMT without order-preservation lexiconized treatment performs well with short sentences and few named entities in a sentence. But as the number of the name entities per sentence increases, especially when the name entities are rare unknowns as discussed before, normal NMT cannot make correct prediction of the right set of name entities with the correct ordering 8. Our lexiconized order-preserving NMT, on the contrary, gives very high accuracy when there are many named entities in the sentence and maintains their correct ordering.

Trimming the lexicon list that keeps the tail helps to increase BLEU scores: Different from most of the previous lexiconized NMT works where BLEU scores never increase (Wang et al., 2017), our BLEU scores show minor improvements. BLEU score for German-Swedish translation increases from 35.8 to 36.6 in Table 7. As an attempt to increase our BLEU scores even further, we conduct two more experiments. In one setting, we keep only the tail of the lexicon table that occur in the Bible once. In another setting, we keep only a manual selection of lexicons. Note that this is the only place where manual work is involved and is not essential. There are minor improvements in BLEU scores in both cases.

33.8% of the translations require minor corrections: The sentence length for these translations that require minor corrections is often longer. We notice that some have repetitions that do not affect meaning, but need to be trimmed. Some have the under-prediction problem where certain named entities in the source sentence never appear; in this case, missing named entities need to be added. Some have minor issues with plurality and tense. We show a few examples of the translations that need minor corrections in the appendices for reference. Typically, sentences with longer sentence length and more complicated named entity relationships require minor corrections to achieve high translation quality.

5 Conclusion and Future Directions

We present our order-preserving translation system for cross-lingual learning in European languages. We examine three issues that are important to translation into low-resource language: the lack of low-resource data, effective cross-lingual transfer, and the variable-binding problem.

Firstly, we add the source and the target family labels in training and examined intra-family and inter-family effects. We find that training on multiple families, more specifically, training on two neighboring families nearest to the low-resource language improves BLEU scores to a reasonably good level. Secondly, we devise a rigorous ablation study and show that we only need a small portion of the low-resource target data to produce reasonably good BLEU scores. Thirdly, to address the variable-binding problem, we build a parallel lexicon table across twenty-three European languages and design a novel order-preserving named entity translation method by tagging named entities in each sentence in order. We achieve reasonably good quantitative and qualitative improvements in a preliminary study.

The order-preserving named entity translation labels named entities in order. Since there are relatively less number of long sentences with many named entities than short sentences with few named entities, underprediction of named entities in long sentences may occur. To seek solution to the underprediction problem, we are looking at randomized labeling of the named entities. Moreover, our order-preserving named entity translation method works well with a fixed pool of named entities in any static document known in advance. This is due to our unique use cases for applications like translating water, sanitation and hygiene (WASH) guidelines written in the introduction. We devise our method to ensure high accuracy targeting translating named entities in static document known in advance. However, researchers may need to translate dynamic document to low-resource language in real-time. We are actively researching into the dynamic timely named entity discovery with high accuracy.

We are actively extending our work to cover more world languages, more diverse domains, and more varied sets of datasets to show our methods are generalizable. Since our experiments shown in this paper are using European languages, we are also interested on non-European languages like Arabic, Indian, Chinese, Indonesian and many others to show that our model is widely generalizable. We also expect to discover interesting research ideas exploring a wider universe of linguistically dissimilar languages.

Our work is helpful for translation into low-resource language, where human translators only need to translate a few lexicons and a partial set of data before passing it to our system. Human translators may also be needed during post-editing before a fully accurate translation is released. Our future goal is to minimize the human correction efforts and to present high quality translation timely.

We would also like to work on real world low-resource tribal languages where there is no or little training data. Translation using limited resources and data in these tribal groups that fits with the culture-specific rules will be very important (Levin et al., 1998). Real world low-resource languages call for cultural-aware translation.

Acknowledgments

We would like to thank Prof. Eduard Hovy for his insights on the topic and helpful suggestions. We would also like to thank Prof. Michael Cysouw for his generous sharing of the massive Bible corpus.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the 17th Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Ulrich Ammon. 2001. *The dominance of English as a language of science: Effects on other languages and language communities*, volume 84. Walter de Gruyter.

Dimitra Anastasiou and Reinhard Schäler. 2010. Translating vital information: Localisation, internationalisation, and globalisation. *Syn-thèses Journal*, 3:11–25.

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.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 21st Conference on Empirical Methods in Natural Language Processing*.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 16th Conference on Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics.

Rafael E Banchs and Marta R Costa-Jussà. 2011. A semantic feature for statistical machine translation. In *Proceedings of the 5th workshop on syntax, semantics and structure in statistical translation*, pages 126–134. Association for Computational Linguistics.

Julia Barrett. 2005. Support and information needs of older and disabled older people in the uk. *Applied ergonomics*, 36(2):177–183.

Stephen Beale, Sergei Nirenburg, Marjorie McShane, and Tod Allman. 2005. Document authoring the bible for minority language translation. *Proceedings of MT-Summit, Phuket, Thailand*.

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.

Sin-wai Chan and David E Pollard. 2001. *An Encyclopaedia of Translation: Chinese-English, English-Chinese*. Chinese University Press.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1693–1703.

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.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1723–1732.

Philipp Dufter and Hinrich Schütze. 2018. A universal semantic space. *arXiv preprint arXiv:1801.06807*.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 949–959.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2017. Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 894–904.

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.

Matthew George Easton. 1897. *Eastons Bible Dictionary: A Dictionary of Bible Terms*. Thomas Nelson.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Low cost portability for statistical machine translation based on n-gram frequency and tf-idf. In *International Workshop on Spoken Language Translation*.

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.

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.

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

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Rosalind M Harding and Robert R Sokal. 1988. Classification of the european language families by genetic distance. *Proceedings of the National Academy of Sciences*, 85(23):9370–9372.

Theo Hermans. 2003. Cross-cultural translation studies as thick translation. *Bulletin of the School of Oriental and African Studies*, 66(3):380–389.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

RD Hitchcock. 1874. Hitchcock's bible names dictionary, art. *AJ Johnson Publishers, New York*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 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.

Lori Levin, Donna Gates, Alon Lavie, and Alex Waibel. 1998. An interlingua based on domain actions for machine translation of task-oriented dialogues. In *Proceedings of the 5th International Conference on Spoken Language Processing*.

Jindřich Libovickỳ and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 196–202.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015. Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, pages 55–60.

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

Martijn Naaijer and Dirk Roorda. 1993. Parallel texts in the hebrew bible, new methods and visualizations. *Young*, 140:157.

Orville James Nave. 1903. *Nave's Topical Bible: A Digest of the Holy Scriptures*. Topical Bible Publishing Company.

Toan Q Nguyen and David Chiang. 2017. Improving lexical choice in neural machine translation. *arXiv preprint arXiv:1710.01329*.

Terence Odlin. 1989. *Language transfer: Cross-linguistic influence in language learning*. Cambridge University Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Robyn Perry and Steven Bird. 2017. Treasure language storytelling: Cross-cultural language recognition and wellbeing. *Proceedings of the 5th International Conference on Language Documentation and Conservation*.

Filippo Petroni and Maurizio Serva. 2008. Language distance and tree reconstruction. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(08):P08012.

Lutz Prechelt. 1998. Early stopping-but when? *Neural Networks: Tricks of the trade*, pages 553–553.

B Reddy, Yadlapalli S Kusuma, Chandrakant S Pandav, Anil Kumar Goswami, Anand Krishnan, et al. 2017. Water and sanitation hygiene practices for under-five children among households of sugali tribe of chittoor district, andhra pradesh, india. *Journal of environmental and public health*.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1-2):129–153.

Edwin W. Rice. 2015. *People's Dictionary of the Bible*. Forgotten Books.

Malcolm Ross et al. 2006. Language families and linguistic diversity. In *Encyclopedia of Language and Linguistics*, 2 edition. Elsevier.

Edward Sapir. 1921. How languages influence each other. *Language: an Introduction to the Study of Speech*.

Kevin P Scannell. 2006. Machine translation for closely related language pairs. In *Proceedings of the Workshop Strategies for developing machine translation for minority languages*, pages 103–109. Citeseer.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Philippa Shoemark, Sharon Goldwater, James Kirby, and Rik Sarkar. 2016. Towards robust cross-linguistic comparisons of phonological networks. In *Proceedings of the 14th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 110–120.

William Smith, Francis Nathan Peloubet, and Mary Abby Thaxter Peloubet. 1967. *Smith's Bible Dictionary*. Pyramid Books.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the 1st Conference on Machine Translation*, volume 2, pages 543–553.

Jörg Tiedemann. 2012. Character-based pivot translation for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 141–151. Association for Computational Linguistics.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? *arXiv preprint arXiv:1801.04962*.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 1357–1366.

Marco Turchi, Tijl De Bie, and Nello Cristianini. 2008. Learning performance of a machine translation system: a statistical and computational analysis. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 35–43. Association for Computational Linguistics.

Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. 2017. Sogou neural machine translation systems for wmt17. In *Proceedings of the 2nd Conference on Machine Translation*, pages 410–415.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Barret Zoph and Kevin Knight. 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.