Complete Multilingual Neural Machine Translation

Markus Freitag and Orhan Firat
Google Research
{freitag,orhanf}@google.com

Abstract

Multilingual Neural Machine Translation (MNMT) models are commonly trained on a joint set of bilingual corpora which is acutely English-centric (i.e. English either as the source or target language). While direct data between two languages that are non-English is explicitly available at times, its use is not common. In this paper, we first take a step back and look at the commonly used bilingual corpora (WMT), and resurface the existence and importance of implicit structure that existed in it: multi-way alignment across examples (the same sentence in more than two languages). We set out to study the use of multi-way aligned examples to enrich the original English-centric parallel corpora. We reintroduce this direct parallel data from multi-way aligned corpora between all source and target languages. By doing so, the English-centric graph expands into a complete graph, every language pair being connected. We call MNMT with such connectivity pattern complete Multilingual Neural Machine Translation (cMNMT) and demonstrate its utility and efficacy with a series of experiments and analysis. In combination with a novel training data sampling strategy that is conditioned on the target language only, cMNMT yields competitive translation quality for all language pairs. We further study the size effect of multi-way aligned data, its transfer learning capabilities and how it eases adding a new language in MNMT. Finally, we stress test cMNMT at scale and demonstrate that we can train a cMNMT model with up to $111*112=12{,}432$ language pairs that provides competitive translation quality for all language pairs.

1 Introduction

Multilingual machine translation (Dong et al., 2015; Firat et al., 2016a; Johnson et al., 2017; Aharoni et al., 2019), which can serve multiple

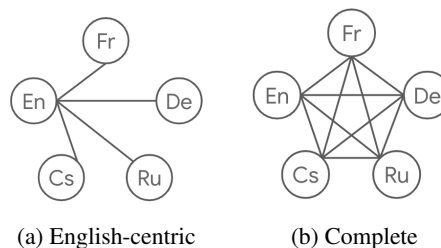(a) English-centric   (b) Complete

Figure 1: Source-target translation graphs in MNMT. Solid lines indicate that there exist direct parallel data. When there is no line connecting any two languages, zero-resource or zero-shot approaches are employed.

language pairs with a single model, has attracted much attention. In contrast to bilingual MT systems which can only serve one single language pair, multilingual models can serve $O(N^2)$ language pairs (N being the number of languages in a multilingual model) (Zhang et al., 2020).

The amount of available training data can differ a lot across language pairs and the majority of available MT training data is English-centric (Tiedemann, 2018; Arivazhagan et al., 2019b) which in practice means that most non-English language pairs do not see a single training example when training multilingual models (see Figure 1a). As a consequence, the actual performance of language pairs that do not include English on the source or target side lags behind the ones with large amounts of training data. Further, when increasing the number of languages, it gets (a) impractical to gather training data for each language pair and (b) challenging to find the right mix during training. Which is why models tasked with direct translation between non-English pairs either resort to bridging (pivoting) through a pivot language (Habash and Hu, 2009), or make use of synthetic parallel data (via back-translation) (Firat et al., 2016b; Chen et al., 2017) or study the problem under zero-shot settings (Johnson et al., 2017; Ha et al., 2016).

In this study, we make use of the potential pre-

550

existing multi-way property in the training corpora and generate as many direct training examples from pre-existing English-centric training data. If we can find training examples for each language pair in a multilingual mix, we call this model complete Multilingual Neural Machine Translation (cMNMT). cMNMT is then trained on all bilingual pairs between source and target languages by utilizing multi-way aligned training examples that consist of translations of the same sentence into multiple languages. We resurface multi-way aligned training examples by aligning training examples from different language pairs when either their source or target sides are identical (ie. pivoting through English, for German→English and English→French to extract German–French–English examples).

To make use of this data, the model samples a source and target language from the set of multi-way aligned corpus during training, which allows the model to see language pairs where originally no training data existed (missing connections in Figure 1a). As our experiments support, this method enables us to get access to training data for all tested language pairs (generating a complete graph (Figure 1b)). We will show that it is possible to generate a complete graph for at least a 6-language WMT setup. Some of the WMT training data is multi-way parallel by construction. Nevertheless, we show that we also find many training examples where the source and target origin from different sources. We further show on our 112 languages internal dataset, that we can find sufficient training data for over 12,000 language pairs by only providing 111 English-centric training corpora. This result indicates that it is possible to generate direct training data for many language pairs without the need for crawling new training examples. Our experiments suggest that before falling back to methods like zero-shot translation, you should investigate the structure of your pre-existing training data.

To address the problem of finding the right mix of examples from different language pairs during training, we further introduce a hierarchical sampling strategy that is language-specific (as opposed to being language pair specific). In addition to fixing some chronic issues of MNMT (i.e. low quality for out of English translation (Firat et al., 2016a; Johnson et al., 2017; Arivazhagan et al., 2019b)), the proposed sampling strategy efficiently ensures all source-target pairs are covered.

Experiments demonstrate that we can train a cMNMT model on a 30-language-pair WMT setup that outperforms bilingual and multilingual baselines as well as bridging on all non-English language pairs. We further show that the performance of the English language pairs stay stable and do not suffer from the changes in both the training data and the new training data sampling strategy. Furthermore, we share experiments at scale by demonstrating that we can train a cMNMT model that can serve 12,432 language pairs.

Our contribution is three-fold:

- We show that we can find a lot of training examples for all language pairs in a multilingual mix by only pivoting pre-existing English-centric training data. We further show that many of the extracted examples originate from different data sources and this method could scale to many more datasets. We also support these findings with experiments on our internal dataset, where we were able to find training data for all 12,432 language pairs.

- We demonstrate that cMNMT outperforms bilingual baselines, multilingual baselines as well as bridging on all non-English language pairs while keeping translation performance on English-centric language pairs.

- We introduce a new sampling strategy that is purely based on the target language instead of language pairs and does scale to MNMT models which hundreds of languages.

2 A Peek at Multi-way Aligned Examples in Bilingual Corpora

We choose six languages Czech (cs), English (en), French (fr), German (de), Spanish (es) and Russian (ru) from the public WMT datasets. The selection of the languages was driven by the fact that the WMT 2013 evaluation campaign (Bojar et al., 2013) released a multi-way test set for these six languages. As training data, we used WMT 2013 for Spanish, WMT 2014 for German, WMT 2015 for French, and WMT 2018 for Czech and Russian.

We can construct non-English bilingual training examples by pairing the non-English sides of two training examples with identical English translations. Table 1 shows the number of bilingual training examples that we could potentially extract from the English-centric training data. The

number of training examples for each non-English language pair varies from at least 0.3 million (Russian-German) to up to 4.8 million sentence pairs (Russian-French).

	cs	de	en	es	fr	ru
cs		0.7	47	0.8	1	0.9
de	0.7		4.5	2.3	2.5	0.3
en	47	4.5		13.1	38.1	33.5
es	0.8	2.3	13.1		10	4.4
fr	1	2.5	38.1	10		4.8
ru	0.9	0.3	33.5	4.4	4.8	

Table 1: WMT: Available training data (in million) after constructing non-English examples from English-centric examples with identical English side.

Some of the extracted non-English training examples are multi-way parallel by construction. The UN corpus is a 6-way parallel corpus, and three of the languages (English, French and Spanish) are in our 6-language mix. A portion of the Europarl corpus is again multi-way aligned. Nevertheless, a good amount of the extracted data is coming from different sources. Table 2 shows the number of non-English bilingual training examples separated by the two sources they originated from.

Table 3 shows how many translations are available for each sentence in the WMT training data. The majority (123 million) of the multi-way aligned examples do only have translations into two languages. As our original bilingual training data is English-centric, all of the 123 million training examples consist of an English sentence and a translation into one of our five other languages. A total of 13 million multi-way aligned examples are available in at least three languages. Further, Figure 2 shows the average number of translations conditioned by the language. Both Spanish and German have, on average more than three translations. In comparison, the majority of the multi-way aligned examples with Czech or English on the target side are bilingual (having only two translations). Our study resurfaced the inherent multi-way aligned information in the commonly used set of parallel corpora instead of discarding this information.

3 Complete Multilingual NMT

We call MNMT models that are trained for all possible source–target pairs as complete MNMT as all languages are connected via training data (also see Figure 1). Before going into details of how
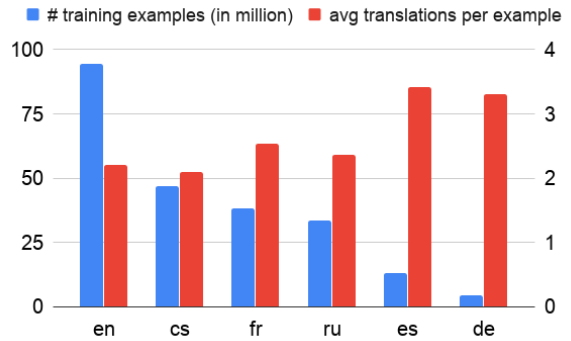

Figure 2: Average translations per multi-way aligned example conditioned on the target language.

the missing pairs' data gathered, we recap MNMT first.

Multilingual NMT Framework   MNMT (Firat et al., 2016b; Johnson et al., 2017) is an extension of bilingual NMT which uses a single model to translate between multiple languages. The model parameters are trained on a joint set of bilingual corpora from different language pairs. Given the data imbalance across the different corpora, it is common to oversample the language pairs with less training data (Lee et al., 2016; Johnson et al., 2017). For a given language pair p, let $D(p)$ be the size of the available parallel corpus, the sample probability with a temperature T is defined as

$$ p_p = \left(\frac{D(p)}{\sum_q D(q)} \right)^{\frac{1}{T}} \tag{1} $$

As a result, $T = 1$ corresponds to the actual data distribution, and $T = 100$ corresponds to (almost) an equal number of samples for each language pair). In addition to being able to translate language pairs that the model was trained with, the model can also translate between language pairs never seen explicitly during training which is often referred as zero-shot translation (Johnson et al., 2017; Ha et al., 2016).

Using multi-way aligned data in MNMT   Instead of only relying on bilingual corpora, bilingual examples from different language pairs with identical target sentences can be combined into a single multi-way aligned training example. An example is given in Table 4. By comparing the English sides of the Spanish–English and the German–English corpora, we extract a multi-way aligned example that contains translations into all three languages.

	cc	CzEng	epps	nc	10^9	Paracrawl	UN	Wiki Titles	Yandex
Common Crawl (cc)	2.5M	13K	213	21k	10k	47k	1.8k	1	6.1k
CzEng 1.7		0	20k	417k	242k	55k	63k	98k	7.7k
Europarl (epps)			6.9M	1.2k	3.7k	4.8k	4.7k	255	280
News-Commentary (nc)				640k	186k	244	305	60	1.7k
10^9					0	12k	97k	1.5k	2.9k
Paracrawl						352k	18k	5.5k	1k
UN							16M	3.7k	3.3k
Wiki Titles								118k	4

Table 2: Number of training examples with identical English sides split by data sources. E.g. cell cc-CzEng shows the number of training examples with identical English side by only considering training data coming from either commoncrawl or CzEng for all language pairs (if available).

# languages	2	3	4	5	6
training data	123M	6.9M	5.4M	0.7M	10k

Table 3: Data statistics for the extracted multi-way aligned training examples for WMT: 123 million sentences are only available in 2 languages, while 10,000 sentences have translations in all 6 languages.

X-Y	*Bleib sicher* ↔ Stay safe
Z-Y	Mantente segura ↔ Stay safe
X-Y-Z	*Bleib sicher* ↔ Mantente segura ↔ Stay safe

Table 4: The two German–English and Spanish–English bilingual training examples can be combined into one multi-way aligned training example that consists of translations into all three languages.

While we can extract direct training data for any source-target pair among the languages considered, the total number of language pairs increases quadratically. The vanilla language pair based sampling strategy in Eq. (1) with adjustable temperature is capable of balancing low-high resource language pairs during training. However, we noticed a critical failure mode, which is further amplified in complete MNMT. The language-pair based sampling strategy (regardless of the temperature being used) *over-represents English* in English-centric models. Notice half of the languages have English on the source side, with the other half on the target side. This over-representation yields a schedule of examples for the encoder (resp. for the decoder) to see English examples half of the time throughout the entire training process. As a result, trained models end up favouring English either on the source and/or target. Although the implications on the encoder could be minimal, over-exposing English examples to the decoder curtail the learning signal when the target language is non-English. We hypothesise that this imbalance in the learning signal with respect to the target language is one of the roots of poor translation quality of multilingual models when translating out of English (Firat et al., 2016a; Johnson et al., 2017; Arivazhagan et al., 2019b).

To alleviate the *over-representation of English* with the language-pair based sampling strategy, we propose a hierarchical sampling strategy with two levels: i) we choose a target language (based on a temperature-based schedule), ii) uniformly sample a source language. Formally, for a given target language l, let $D(l)$ be the size of the available training examples with target language l, the sample probability with a temperature T is defined as

$$p_l = \left(\frac{D(l)}{\sum_q D(q)} \right)^{\frac{1}{T}} \qquad (2)$$

During training, the scheduler samples a batch of training examples based on the target language only, as opposed to source-target language pair specific sampling. After choosing a target language, for each multi-way aligned example, we randomly (uniformly) pick one of the translations as the source sentence.

4 Experiments

We use a public transformer implementation with the transformer-big model size (Vaswani et al., 2017) for all multilingual setups. All bilingual models use a vocabulary of 32,000 subwords, while all multilingual models use a vocabulary of 64,000 subword units. All multilingual models are trained for 500,000 updates using an average batch size of around 33,000 sentences (∼1 million tokens). All bilingual models are trained for 400,000 steps as they converged earlier using a batch size of around 8,000 sentences (∼260,000 tokens). Due to the data imbalance across languages, we use a

temperature-based data sampling strategy to over-sample low-resource language pairs in standard MNMT models (Equation 1) and low-resource target languages in cMNMT models (Equation 2). We use a temperature of $T = 5$ in both cases. All multilingual models add a token at the beginning of the input sentence to specify the required target language. All BLEU (Papineni et al., 2002) scores are calculated with sacreBLEU (Post, 2018).[1]

4.1 Baselines on WMT

We train several baselines: (i) bilingual models, (ii) multilingual models based on English-centric data, and (iii) bridging non-English language pairs.

Bilingual Baselines We train two bilingual baselines (using either transformer-base or transformer-big) for each language pair. In addition to training baselines on the original English-centric WMT data, we also train models for non-English language pairs on the extracted direct data (see Table 1). We experimented with several dropout rates for both setups and found that dropout=0.1 works best for transformer-base while dropout=0.3 works best for transformer-big. As can be seen from Table 5 and Table 6, the experiments suggest that the translation quality of the non-English language pairs is far behind the ones for English-centric language pairs. As an example, the translation quality between German and Russian reaches 6-7 BLEU only.

	target					
source		cs	de	en	es	fr	ru
cs			16.6	30.4	20.7	22.6	13.9
de		15.2		29.5	27.0	28.7	6.9
en		25.2	25.7		33.6	34.8	23.6
es		15.6	22.7	33.9		34.2	18.7
fr		15.4	22.1	33.0	31.8		17.9
ru		12.5	6.1	28.4	22.6	24.3	

Table 5: BLEU scores on newstest2013 of bilingual models trained with the transformer-base architecture.

Multilingual Baselines We train a multilingual NMT model on the original WMT English-centric training data. BLEU scores are summarized in Table 7. All language pairs with English as the source or target language perform comparably well from at least 24.5 BLEU (English→Russian) up to 34.9 BLEU (English→French). The BLEU scores of

	target					
source		cs	de	en	es	fr	ru
cs			14.6	31.9	19.0	20.0	14.1
de		14.1		31.3	26.4	28.8	4.7
en		26.5	27.0		34.2	35.9	25.0
es		14.4	22.8	34.5		34.8	19.9
fr		13.0	20.7	34.2	32.5		18.6
ru		12.8	4.0	30.8	23.1	24.8	

Table 6: BLEU scores on newstest2013 of bilingual models trained with the transformer-big architecture.

non-English language pairs are consistently lower (which can be explained as a lack of supervision during training) and can be as low as 4.1 BLEU for Spanish→Czech or as high as 24.4 BLEU for French→Spanish.

	target					
source		cs	de	en	es	fr	ru
cs			19.8	31.2	21.6	20.2	8.5
de		6.8		31.8	17.8	21.2	4.5
en		25.5	26.7		34.0	34.9	24.5
es		4.1	8.8	34.7		19.6	9.5
fr		4.2	11.2	33.8	24.4		6.5
ru		4.8	10.4	29.5	19.9	9.6	

Table 7: BLEU scores on newstest2013 of a MNMT model trained on English-centric training data. All non-English language pairs are unseen during training and BLEU scores measure zero-shot performance.

Bridging (Pivoting) Baselines The quality of MNMT is still behind the one from bilingual baselines for most of the language pairs (comparing Table 6 and Table 7). Nevertheless, having a single NMT model for each language pair is impractical, especially when increasing the number of language pairs. An alternative approach is called bridging (Cohn and Lapata, 2007; Wu and Wang, 2007; Utiyama and Isahara, 2007). For the bridging approach, we compromise and train only English-centric models. To enable the translation between non-English language pairs, the source sentence cascades through the source→English and English→target systems to generate the target sentence. This simple process has several limitations: (i) translation errors accumulate in the pipeline, (ii) decoding time gets doubled since inference has to be run twice, (iii) bridging through a morphologically low language (i.e. English), important information could be lost (i.e. gender). The BLEU scores (Table 8) for all non-English pairs are higher compared to all previous baselines. We can reach acceptable translation quality even for

---

[1] sacreBLEU signatures: BLEU+case.mixed+lang.SRC-TGT+numrefs.1+smooth.exp+SET+tok.intl+version.1.2.20

German→Russian, where our direct training data is scarce. We use the bridging baseline to compare our cMNMT models in the rest of the paper.

	target					
source	cs	de	en	es	fr	ru
cs		22.4	31.9	27.0	28.8	21.9
de	21.5		31.3	26.9	29.0	20.3
en	26.5	27.0		34.2	35.9	24.9
es	22.6	22.8	34.5		32.6	22.5
fr	21.4	22.2	34.2	29.1		21.6
ru	21.3	20.6	30.8	27.2	28.5	

Table 8: BLEU scores on newstest2013 for our WMT setup. Translations for non-English language pairs are generated via bridging over English.

4.2 Complete MNMT Models on WMT

Without adding new training data and taking into account the multi-way property of the data, we train a complete multilingual NMT system (cMNMT, see Section 3). We compare the performance of cMNMT with the best baseline model that is based on bridging (Table 8) and report BLEU and delta BLEU numbers in Table 9. The BLEU scores for the non-English language pairs go up from at least 1.4 BLEU for Russian→Spanish up to 5.0 BLEU for Czech→Russian. We changed the sampling strategy for our cMNMT models to be conditioned on the target language only (Section 3). As a result, English has been seen less often as the target language when compared to a standard MNMT setup. Interestingly, this seems to affect only the performance of Russian→English, which shows a decrease of 1 BLEU point. The other language pairs with English as the target language are keeping their translation quality.

When comparing our cMNMT model to the English-centric baseline (Table 7), we see an average BLEU increase of 14.6 BLEU for all non-English language pairs. It is worth noticing that every language pair has now at least 22 absolute BLEU points. Interestingly, the absolute BLEU scores in each row (translations into the same language) are much closer, suggesting a more universal input representation.

5 Analysis and Discussion

To further understand the impact of multi-way aligned examples on NMT, we run a couple of additional experiments.

	target					
source	cs	de	en	es	fr	ru
cs		25.8 +3.4	32.0 +0.1	30.1 +3.1	31.4 +2.6	26.9 +5.0
de	23.9 +2.4		31.2 -0.1	29.9 +3.0	31.8 +2.8	23.4 +3.1
en	26.9 +0.4	27.1 +0.1		35.0 +0.8	35.5 -0.4	26.4 +1.5
es	24.9 +2.3	25.7 +2.9	34.9 +0.4		36.0 +3.4	24.9 +2.4
fr	23.7 +2.3	25.2 +3.0	34.2 +0.0	33.3 +4.2		23.5 +1.9
ru	24.3 +3.0	22.7 +2.1	29.8 -1.0	28.6 +1.4	30.1 +1.6	

Table 9: BLEU on newstest2013 for our novel cMNMT model. The small numbers are the difference (ΔBLEU) with respect to the bridging approach (Table 8).

Training Data Sampling Strategy In Section 3, we did introduce our new training data sampling strategy that is based on the target language only. This change was mainly driven by the fact that having a language-pair conditioned schedule is not scalable when building a system of 12,432 language pairs. Instead of finding a good sampling weight for each of the 12,432 language pairs, we only need to find a suitable mix for the 112 target languages. Further, we have more control over how often each target language will be seen during training. To see the impact of this change, we train an MNMT system on the joint set of the 30 different bilingual corpora with a standard language-pair based temperature scheduling scheme and compare it to a cMNMT model. We used temperature 5 in both setups. ΔBLEU numbers for each language-pair can be seen in Figure 3. The language-conditioned temperature scheduling increases BLEU scores for 29 out of 30 language-pairs with larger gains for the low-resource language-pairs. This experiment suggests that a target language based temperature scheduling is not only simpler but also performs better on average.

Separate Multi-way Aligned Examples We test the transfer learning capability of cMNMT by training a cMNMT model only on the 13 million multi-way aligned examples that have translations in at least three languages (see Table 3). In other words, we remove all training examples that are only available in English and one additional language. If no transfer learning is happening, the English-centric scores will decrease while the BLEU numbers of the non-English language
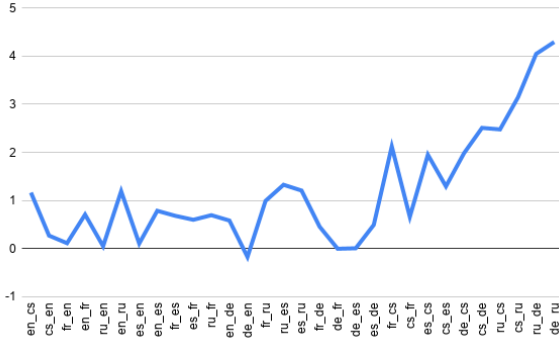
Figure 3: ΔBLEU scores for a target language based versus a language-pair based temperature schedule.

pairs are not affected. Experimental results can be seen in Table 10. Interestingly, we find that the performance of all language pairs is similarly affected. This indicates that transfer learning is happening between the language pairs and that non-English language pairs benefit from having more English-centric data.

	target					
source	cs	de	en	es	fr	ru
cs		24.1 -1.7	30.3 -1.7	28.4 -1.7	29.4 -2.0	23.6 -3.3
de	19.1 -4.8		30.6 -0.6	29.1 -0.8	30.5 -1.3	21.7 -1.7
en	21.8 -5.1	26.2 -0.9		33.4 -1.6	34.6 -0.9	23.3 -3.1
es	19.7 -5.2	24.9 -0.8	34.3 -0.6		35.1 -0.9	22.7 -2.2
fr	18.6 -5.1	24.0 -1.2	33.0 -1.2	32.3 -1.0		21.3 -2.2
ru	19.7 -4.6	21.4 -1.3	27.7 -1.9	27.1 -1.5	27.9 -2.2	

Table 10: BLEU on newstest2013 for a model trained on 13 million multi-way aligned ($n>2$) data. Small numbers are the difference (ΔBLEU) between cMNMT trained on all multi-way examples (136M, Table 9).

To further study this effect, we reverse that experiment and remove all examples that have translations into more than two languages. This experiment investigates if the non-English language pairs in a standard MNMT model can benefit from having training examples with identical English sides. Experimental results can be found in Table 11. The BLEU scores for English-centric language pairs drop by 0.9 points on average while the performance of non-English language pairs decreases by 1.6 BLEU on average.

	target					
source	cs	de	en	es	fr	ru
cs		17.8 -2.0	31.5 +0.3	14.5 -7.1	20.3 +0.1	5.2 -3.3
de	7.8 +1.0		29.6 -2.2	17.2 -0.6	22.9 +1.7	1.8 -2.7
en	25.6 +0.1	23.9 -2.8		33.1 -0.9	33.5 -1.4	24.6 +0.1
es	7.2 +3.0	3.9 -5.8	32.7 -1.1		24.3 +4.7	7.6 -1.9
fr	6.7 -2.5	13.0 -1.8	33.1 -0.7	19.3 -5.1		7.2 +0.7
ru	5.5 +0.7	10.4 +0.0	29.3 -0.2	8.5 -14.4	13.7 +4.1	

Table 11: BLEU on newstest2013 for a model trained on 2-way data only. Small numbers are the difference (ΔBLEU) between the vanilla MNMT model (Table 7).

Leave N-Out We further investigate the transfer learning capability of our approach by training several cMNMT models on different amounts of training data. We start with a cMNMT model trained on English-centric bilingual training data only. This setup ensures that all languages have been seen on both the source and target side during training. We further group the remaining multi-way aligned training examples by target language and add one after another to the training data. Important to mention: We retrained all configurations from scratch. Experimental results are summarized in Figure 4. We report average BLEU scores grouped by the target language. We can see that adding training data $x \rightarrow y$ for a target language y, gives a significant boost in translation quality for that target language. These results demonstrate that even though we can translate between language pairs without seeing a single example during training, adding supervision during training significantly increases BLEU scores.

Adding a New Language We further investigate how a cMNMT model behaves when fine-tuned (Freitag and Al-Onaizan, 2016) to a new language. We chose Italian as the new language as the test set newstest2009 is multi-way in Czech, English, French, German, Italian and Spanish and thus we can report BLEU scores between all language pairs. We run two experiments with two different sets of fine-tuning data. First, we fine-tuned the cMNMT model (Table 9) on English↔Italian news-commentary (45,000 examples). Second, we converted the same data into multi-way aligned examples by augmenting the bilingual examples with translations into other languages when found in our
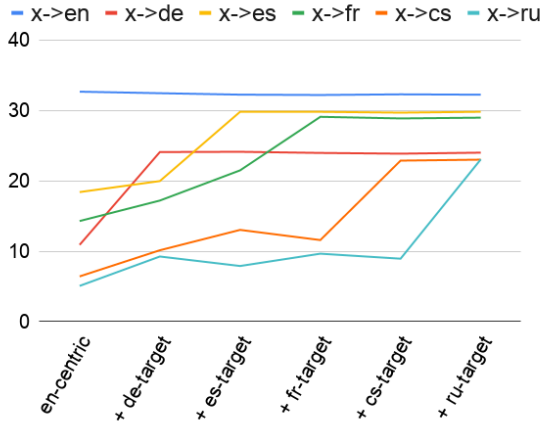
556

Figure 4: BLEU scores of models using only parts of the multi-way training data.

Figure 5: Average translations per multi-way example conditioned on the target language.

Arivazhagan et al. (2019b) for more details regarding the dataset. Figure 5 shows the training data sizes and the average translations per multi-way example.
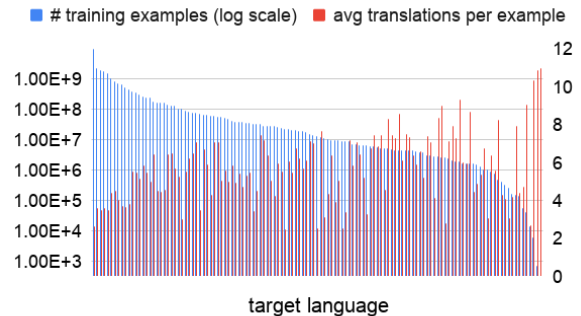
Although a deeper and wider architecture does improve the quality of multilingual models for this dataset, we use the same experimental setup as used in our WMT experiments (see Section 4) to run an MNMT and cMNMT model on our in-house data. Experimental results can be seen in Table 13. cMNMT outperforms MNMT for non-English languages by 10.1 BLEU points on average while keeping the translation quality for language pairs that include English as source or target. These results demonstrate that our proposed approach does scale far behind the six language WMT setup.

original training data. Experimental results for fine-tuning our model for one epoch on either of the two datasets can be found in Table 12. Both fine-tuning experiments show the same BLEU improvements for Italian↔English. Nevertheless, when only fine-tuning on English↔Italian data, we sacrifice translation quality for most of the language pairs which can be seen in the $x \to y$ column. Further, fine-tuning on multi-way aligned examples does improve the average BLEU scores by 4.3 BLEU for translations into Italian ($x \to$it). Overall, these experiments suggest that fine-tuning with multi-way aligned data is superior.

model	it→en	it→ x	en→it	$x \to$it	$x \to y$
cMNMT	13.5	9.7	2.3	2.6	22.0
+ft en↔it	21.5	14.2	13.6	11.8	17.8
+ft mway	21.2	18.5	13.5	11.9	23.0

Table 12: BLEU scores for newstest2009 for fine-tuning (ft) our cMNMT model on either English↔Italian (it↔en) news-commentary or on the same sentences but augmented with translations into other languages (mway), if available. Column $x \to y$ shows average BLEU scores for all language pairs.

Scaling cMNMT: 12,432 Language Pairs We run additional experiments on a 112 language in-house dataset (Arivazhagan et al., 2019b) to see if our approach scales to 12,432 language pairs. Our in-house dataset does not only contain more languages than the WMT setup, but also has a much wider range of available training resources. While for the high resource languages, we have access to billions of training examples, most of the low resource languages have less than 1 million training examples. We refer the reader to the description in

	en→ x	$x \to$en	$x \to y$
all	+0.34	-0.05	+10.1
low resource	+0.23	-0.15	+8.82
mid resource	+0.35	-0.05	+11.02
high resource	+0.45	+0.04	+9.73

Table 13: Average BLEU difference (ΔBLEU) between a cMNMT and a vanilla MNMT model for our in-house 112 language setup. Positive numbers present improvement of cMNMT over MNMT.

6 Related Work

Direct models To translate between languages with little training data, three general approaches emerged, i. bridging through a third language (pivot-based MT) (Cheng et al., 2016; Currey and Heafield, 2019), ii. generating pseudo-parallel data between direct language pairs and training the direct pairs with that (zero-resource MT) (Firat et al., 2016b; Chen et al., 2017) and, iii. zero-shot methods where the model is asked to translate a direct pair only at test time (Johnson et al., 2017; Ha et al., 2016; Arivazhagan et al., 2019a).

Although pivot-based approaches perform sufficiently good when cascaded with strong bilingual models (Gu et al., 2019), their practicality is limited due to compounding errors from pipelining and doubled inference cost. The zero-resource approaches, combined with iterative-back translation (Hoang et al., 2018) are quite powerful but their inefficiency is worth noting. For N languages, one needs to devise a training routine that could sample $N^2 - N$ pairs, generate pseudo-parallel data. The added time to generate pseudo-parallel data for every pair grows quadratically, making it challenging for systems considering a large number of languages. Recently, by devising a practical sub-sampling approach, (Zhang et al., 2020) demonstrated zero-resource techniques could be scaled to massively multilingual setup. We find the study by (Zhang et al., 2020) closest to our work, having the goal of any-to-any multilingual translation. But compared to sampling language pairs with no parallel data and generating pseudo-parallel data on-the-fly, our approach makes use of existing multi-way alignment information before training. Lastly, zero-shot approaches attempt to measure the generalization performance of the MNMT models, but to date, the zero-shot quality still trails behind the pivot and zero-resource methods (Al-Shedivat and Parikh, 2019). Our proposed cMNMT, naturally fills the gap between these three approaches, the multi-way data can be extracted offline, and efficiently be mixed with the original data using a hierarchical data sampler. It does not require extra steps to generate pseudo-parallel data, and (as expected) it handily outperforms zero-shot approaches.

N-way data In this paper, we only made use of multi-way aligned data to sample bilingual pairs out of it. But there exist several approaches that make use of the multi-view structure in the data, such as Dabre et al. (2019), who explored the use of small multi-parallel corpora a for one-to-many NMT. Another approach is multi-source NMT (Zoph and Knight, 2016). Although multi-source NMT is a promising direction, it has practical problems such as lacking multiple sources at inference time (Nishimura et al., 2018). We believe research in this direction will be the key to improve mid/high-resource NMT and address several robustness issues to the input noise. Aulamo et al. (2020) recently released MultiParaCrawl where the authors extracted direct data for non-English lan-

guage pairs from the English-centric Paracrawl corpus.

Sampling scheduling Several approaches proposed to address data sampling for multi-task models, some relying on temperature-based heuristics (Lee et al., 2016; Devlin et al., 2018; Arivazhagan et al., 2019b), others relying on adaptive schedules that incorporate the model gains, baselines or quality expectations into the data schedulers (Kiperwasser and Ballesteros, 2018; Jean et al., 2019; Wang et al., 2020). We believe data sampling is a critical research area for not only MNMT but also multi-task learning in general. We reveal a critical failure mode of the commonly used temperature sampling strategy, and how it causes the poor translation quality while translating out of English.

7 Conclusion

In this work, we introduced complete Multilingual Neural Machine Translation (cMNMT) that exploits the multi-way alignment information in the underlying training data to improve translation quality for language pairs where training data is scared or not available. Standard MNMT models are trained on a joint set of different training corpora for a variety of language pairs. cMNMT combines the different corpora and constructs multi-way aligned training examples that consist of translations of the same sentence into multiple languages. In combination with a novel temperature-based sampling approach that is conditioned on the target language only, we show that cMNMT is superior to the standard MNMT model and the even better-performing bridging approach. Experimental results on a public WMT 30 language pairs dataset and an in-house 12,432 language pairs dataset demonstrated an average BLEU increase of more than 10 BLEU points for non-English language pairs. This approach leads to a single NMT model that can serve 12,432k language pairs with reasonable quality which also surpasses the translation quality of the bridging approach, which is nowadays used in most modern MT services.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively Multilingual Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

Volume 1 (Long and Short Papers), pages 3874–3884.

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1184–1197.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019a. The missing ingredient in zero-shot neural machine translation. *CoRR*, abs/1903.07091.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019b. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, and Jörg Tiedemann. 2020. Opustools and parallel corpus diagnostics. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3782–3789.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Yun Chen, Yang Liu, Yong Cheng, and Victor OK Li. 2017. A Teacher-Student Framework for Zero-Resource Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935.

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2016. Neural machine translation with pivot languages. *CoRR*, abs/1611.04928.

Trevor Cohn and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic. Association for Computational Linguistics.

Anna Currey and Kenneth Heafield. 2019. Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107.

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation.

In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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 and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. *CoRR*, abs/1906.01181.

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

Nizar Habash and Jun Hu. 2009. Improving Arabic-Chinese statistical machine translation using English as pivot language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181, Athens, Greece. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Sébastien Jean, Orhan Firat, and Melvin Johnson. 2019. Adaptive scheduling for multi-task learning.

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.

Eliyahu Kiperwasser and Miguel Ballesteros. 2018. Scheduled multi-task learning: From syntax to translation. *Transactions of the Association for Computational Linguistics*, 6:225–240.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *CoRR*, abs/1610.03017.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018. Multi-source neural machine translation with missing data. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 92–99, Melbourne, Australia. Association for Computational Linguistics.

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 of the Association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting Bleu Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Jörg Tiedemann. 2018. Emerging language spaces learned from massively multilingual corpora. *CoRR*, abs/1802.00273.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.