Evaluating the Supervised and Zero-shot Performance of Multi-lingual Translation Models

Chris Hokamp and **John Glover** and **Demian Gholipour**
Aylien Ltd.
Dublin, Ireland
`<first-name>@aylien.com`

Abstract

We study several methods for full or partial sharing of the decoder parameters of multilingual NMT models. Using only the WMT 2019 shared task parallel datasets for training, we evaluate both fully supervised and zero-shot translation performance in 110 unique translation directions. We use additional test sets and re-purpose evaluation methods recently used for unsupervised MT in order to evaluate zero-shot translation performance for language pairs where no gold-standard parallel data is available. To our knowledge, this is the largest evaluation of multi-lingual translation yet conducted in terms of the total size of the training data we use, and in terms of the number of zero-shot translation pairs we evaluate. We conduct an in-depth evaluation of the translation performance of different models, highlighting the trade-offs between methods of sharing decoder parameters. We find that models which have task-specific decoder parameters outperform models where decoder parameters are fully shared across all tasks.

1 Introduction

Multi-lingual translation models, which can map from multiple source languages into multiple target languages, have recently received significant attention because of the potential for positive transfer between high- and low-resource language pairs, and because of the potential efficiency gains enabled by translation models which share parameters across many languages (Dong et al., 2015; Ha et al., 2016; Firat et al., 2016; Johnson et al., 2016; Blackwood et al., 2018; Sachan and Neubig, 2018; Aharoni et al., 2019). Multi-lingual models which share parameters across tasks can also perform zero-shot translation, translating between language pairs for which no parallel training data is available (Wu et al., 2016; Ha et al., 2016; Johnson et al., 2016).

Although multi-task models have recently been shown to achieve positive transfer for some combinations of NLP tasks, in the context of MT, multi-lingual models do not universally outperform models trained to translate in a single direction when sufficient training data is available. However, the ability to do zero-shot translation may be of practical importance in many cases, as parallel training data is not available for most language pairs (Wu et al., 2016; Johnson et al., 2016; Aharoni et al., 2019). Therefore, small decreases in the performance of supervised pairs may be admissible if the corresponding gain in zero-shot performance is large. In addition, zero-shot translation can be used to generate synthetic training data for low- or zero- resource language pairs, making it a practical alternative to the bootstrapping by back-translation approach that has recently been used to build completely unsupervised MT systems (Firat et al., 2016; Artetxe et al., 2018; Lample et al., 2018a,b). Therefore, understanding the trade-offs between different methods of constructing multi-lingual MT systems is still an important line of research.

Deep sequence to sequence models have become the established state-of-the-art for machine translation. The dominant paradigm continues to be models divided into roughly three high-level components: *embeddings*, which map discrete tokens into real-valued vectors, *encoders*, which map sequences of vectors into an intermediate representation, and *decoders*, which use the representation from an encoder, combined with a dynamic representation of the current state, and output a sequence of tokens in the target language conditioned upon the encoder's representation of the input. For multi-lingual systems, any combination of encoder and/or decoder parameters can potentially be shared by groups of tasks, or duplicated and kept private for each task.
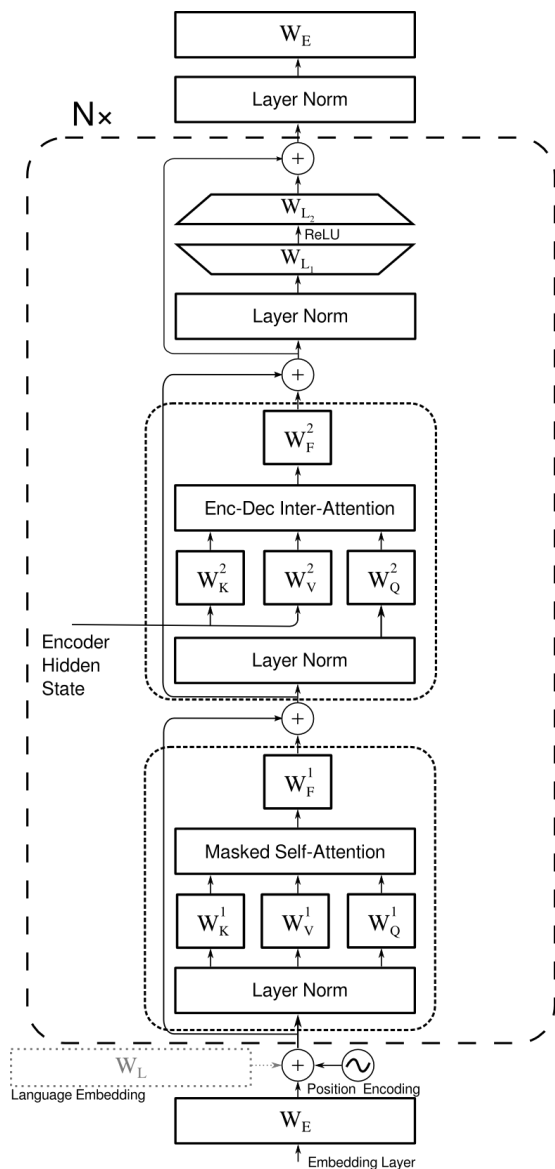
Figure 1: The decoder component of the transformer model (Vaswani et al., 2017). We can share all parameters across all target tasks, or we can create a unique set of decoder parameters for each task (outer dashed line). Alternatively, we can create unique attention parameters for each task, while sharing the final feed-forward layers (inner dotted lines). The possiblility of including an embedding for the target task is visualized at the bottom of the diagram. Illustration modeled after Sachan and Neubig (2018).

Our work builds upon recent research on many-to-one, one-to-many, and many-to-many translation models. We are interested in evaluating many-to-many models under realistic conditions, including:

1. A highly imbalanced amount of training data available for different language pairs.

2. A very diverse set of source and target languages.

3. Training and evaluation data from many domains.

We focus on multi-layer transformer models (Vaswani et al., 2017), which achieve state-of-the-art performance on large scale MT and NLP tasks (Devlin et al., 2018; Bojar et al., 2018). We study four ways of building multi-lingual translation models. Importantly, all of the models we study can do zero-shot translation: translating between language pairs for which no parallel data was seen at training time. The models use training data from 11 distinct languages[1], with supervised data available from the WMT19 news-translation task for 22 of the 110 unique translation directions[2]. This leaves 88 translation directions for which no parallel data is available. We try to evaluate zero-shot translation performance on all of these additional directions.

Target Language Specification Although the embedding and encoder parameters of a multi-lingual system may be shared across all languages without any special modification to the model, *decoding* from a multi-lingual model requires a means of specifying the desired output language. Previous work has accomplished this in different ways:

- pre-pending a special target-language token to the input (Wu et al., 2016)

- using an additional embedding vector for the target language (Lample and Conneau, 2019)

- using unique decoders for each target language (Luong et al., 2016; Firat et al., 2016)

- partially sharing some of the decoder parameters while keeping others unique to each target language (Sachan and Neubig, 2018; Blackwood et al., 2018)

However, to the best of our knowledge, no side-by-side comparison of these approaches has been conducted. We therefore train models which are identical except for the way that decoding into different target languages is handled, and conduct

---

[1] CS, DE, EN, FI, FR, GU, KK, LT, RU, TR and ZH
[2] Note we do not consider auto-encoding, thus the number of translation directions is $11^2 - 11 = 110$.

a large-scale evaluation. We use only the language pairs and official parallel data released by the WMT task organisers, meaning that all of our systems correspond to the constrained setting of the WMT shared task, and our experimental settings should thus be straightforward to replicate.

2 Multi-Task Translation Models

This section discusses the key components of the transformer-based NMT model, focusing on the various ways to enable translation into many target languages. We use the terms source/target *task* and *language* interchangeably, to emphasize our view that multi-lingual NMT is one instantiation of the more general case of multi-task sequence to sequence learning.

2.1 Shared Encoders and Embeddings

In this work, we are only interested in ways of providing target task information to the model – information about the source task is never given explicitly, and the encoder is always completely shared across all tasks. The segmentation model and embedding parameters are also shared between all source and target tasks (see below for more details).

2.2 Multi-lingual Decoder Configurations

Figure 1 visualizes the decoder component of the transformer model, with dashed and dotted lines indicating the parameter sets that we can replicate or share across target tasks.

2.2.1 Target Task Tokens (PREPEND)

Wu et al. (2016) showed that, as long as a mechanism exists for specifying the target task, it is possible to share the decoder module's parameters across all tasks. In the case where all parameters are shared, the decoder model must therefore learn to operate in a number of distinct modes which are triggered by some variation in the input. A simple way to achive this variation is by pre-pending a special "task-token" to each input. We refer to this method as **PREPEND**.

2.2.2 Task Embeddings (EMB)

An alternative to the use of a special task token is to treat the target task as an additional input feature, and to train a unique embedding for each target task (Lample and Conneau, 2019), which is combined with the source input. This technique has the advantage of explicitly decoupling target

task information from source task input, introducing a relatively small number of additional parameters. This approach can be seen as adding an additional token-level *feature* which is the same for all tokens in a sequence (Sennrich and Haddow, 2016). We refer to this setting as **EMB**.

2.2.3 Task-specific Decoders (DEC)

In general, any subset of decoder parameters may be replicated for each target language, resulting in parameter sets which are specific to each target task. At one extreme, the entire decoder module may be replicated for each target language, a setting which we label **DEC** (Dong et al., 2015).

2.2.4 Task-specific Attention (ATTN)

An approach somewhere in-between EMB and DEC is to partially share some of the decoder parameters, while keeping others unique to each task. Recent work proposed creating unique attention modules for every target task, while sharing the other decoder parameters (Sachan and Neubig, 2018; Blackwood et al., 2018). The implementation of their approaches differ significantly – we propose to create completely unique attention parameters for each task. This means that for each of our 11 languages, we have unique context- and self-attention parameters in each layer of the transformer decoder. We refer to this setting as **ATTN**.

3 Experiments

All experiments are conducted using the transformer-base configuration of Vaswani et al. (2017) with the relevant modifications for each system discussed in the previous section. We use a shared sentencepiece[3] segmentation model with 32000 pieces. We use all available parallel data from the WMT19 news-translation task for training, with the exception of `commoncrawl`, which we found to be very noisy after manually checking a sample of the data, and `paracrawl`, which we use only for EN-FI and EN-LT[4].

We train each model on two P100 GPUs with an individual batch size of up to 2048 tokens. Gradients are accumulated over 8 mini-batches and parameters are updated synchronously, meaning that our effective batch size is $2 * 2048 * 4 = 16384$ tokens per iteration. Because the task pair for

---

[3]https://github.com/google/sentencepiece
[4]Turkish (TR) is included from the 2018 language pairs because the task-organizers suggest the possibility of using TR data to improve KK performance

	# seen	# available	# epochs	% budget
EN-CS	3,466,692	51,136,198	0.06	10.7
EN-DE	2,678,808	3,054,632	0.88	8.3
EN-FI	3,466,692	6,457,071	0.54	10.7
EN-GU	1,260,615	137,905	9.14	3.9
EN-KK	1,181,827	158,067	7.47	3.7
EN-LT	3,624,269	2,283,272	1.59	11.2
EN-RU	5,042,462	11,391,126	0.44	15.6
EN-TR	1,575,769	207,678	7.58	4.9
EN-ZH	5,846,104	14,549,833	0.40	18.1
DE-FR	4,097,000	1,980,332	2.06	12.7
TOTAL	32,240,238	91,356,114	-	100

Table 1: Training dataset statistics for our multilingual NMT experiments. **# seen** is the total number of segments seen during training. **# available** is the number of unique segments available in the parallel training datasets. **# epochs** is the number of passes made over the available training data – when this is < 1, the available training data was only partially seen. **% budget** is the percentage of the training budget allocated to this pair of tasks.

each mini-batch is sampled according to our policy weights and (fixed) random seed, and each iteration consists of 8 unique mini-batches, a single parameter update can potentially contain information from up to 8 unique task pairs. We train each model for 100,000 iterations without early stopping, which takes about 40 hours per model. When evaluating we always use the final model checkpoint (i.e. the model parameters saved after 100,000 iterations). We use our in-house research NMT system, which is heavily based upon OpenNMT-py (Klein et al., 2017).

The sampling policy weights were specified manually by looking at the amount of available data for each pair, and estimating the difficulty of each translation direction. The result of the sampling policy is that lower resource language pairs are upsampled significantly. Table 1 summarizes the statistics for each language pair. Note that the data in each row represents a *pair* of tasks, i.e. the total number of segments seen for EN-CS is split evenly between EN→CS, and CS→EN. Because we train for only 100,000 iterations, we do not see all of the available training data for some high-resource language pairs.

With the exception of the system which prepends a target task token to each input, the input to each model is identical. Each experimental setting is mutually exclusive, i.e. in the EMB setting we do not prepend task tokens, and in the ATTN setting we do not use task embeddings.

Figure 2 plots the validation performance during training on one of our validation datasets. The language embeddings from the EMB system are visualized in figure 3.
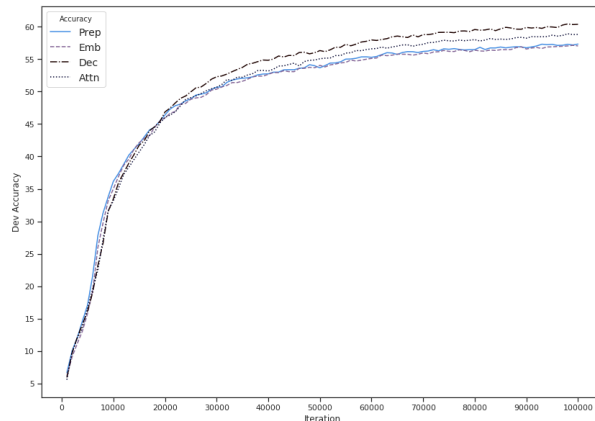
3.1 Results

Figure 2: Word-level accuracy on WMT EN-DE 2014 dev set as training progresses. The model which has a DE-specific decoder achieves the highest accuracy on this dev set.
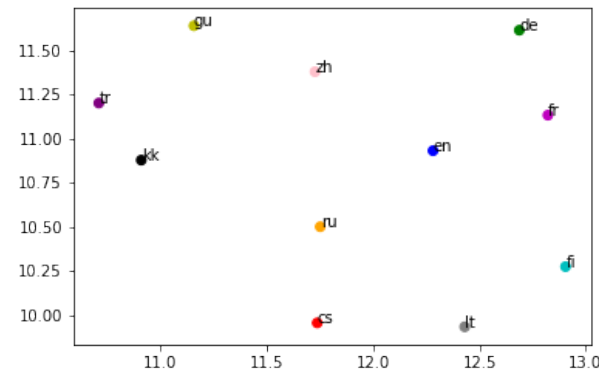
Figure 3: Language embeddings of the EMB system projected with UMAP (McInnes et al., 2018).

We evaluate the performance of our models in four ways. First, we check performance on the supervised pairs using dev and test sets from the WMT shared task. We then try to evaluate zero-shot translation performance in several ways. We use the TED talks multi-parallel dataset (Ye et al., 2018) to create gold sets for all zero-shot pairs that occur in the TED talks corpus, and evaluate on those pairs. We also try two ways of evaluating zero-shot translation without gold data. In the first, we do round-trip translation SRC $\rightarrow$ PIVOT $\rightarrow$ $\widehat{\text{SRC}}$, and measure performance on the $(\widehat{\text{SRC}}, \text{SRC})$ pair – this method is labeled

	Evaluation Dataset
EN-CS	newstest2018
EN-DE	newstest2018
EN-FI	newstest2018
EN-GU	newsdev2019
EN-KK	newsdev2019
EN-LT	newsdev2019
EN-RU	newstest2018
EN-TR	newstest2018
EN-ZH	newstest2018
DE-FR	euelections_dev2019

Table 2: The WMT evaluation dataset used for each language pair.

	PREPEND	EMB	DEC	ATTN
SUPERVISED	23.4	23.4	24.0	**24.1**
ZERO-SHOT-TED	10.6	7.8	**12.6**	12.4
ZERO-SHOT-PIVOT	16.9	**18.1**	14.0	15.1
ZERO-SHOT-PARALLEL-PIVOT	13.1	11.9	12.8	**13.2**

Table 3: Overall results for supervised and zero-shot tasks. Tokenized BLEU scores are computed by concatenating all of the hypotheses for all translation directions, and computing BLEU with respect to the concatenated references. We use the sentencepiece-segmented hypotheses and references to avoid issues with tokenization of multi-lingual hypotheses and references.

	PREPEND	EMB	DEC	ATTN
CS-EN	20.2	20.2	**20.9**	20.9
EN-CS	12.4	12.7	**13.7**	13.3
DE-EN	26.2	26.1	**27.4**	27.1
EN-DE	23.2	23.4	**25.7**	25.2
FI-EN	13.7	13.5	**14.4**	14.2
EN-FI	8.3	8.0	**9.4**	9.2
GU-EN	15.4	15.4	**15.7**	15.4
EN-GU	**8.1**	7.8	5.1	7.3
KK-EN	**14.4**	14.0	14.3	13.9
EN-KK	**5.6**	5.2	1.9	4.6
LT-EN	18.6	18.9	**19.3**	19.0
EN-LT	12.8	13.0	**14.4**	13.7
RU-EN	20.8	20.6	21.3	**21.3**
EN-RU	15.5	15.9	**17.0**	16.7
TR-EN	14.8	15.0	**15.2**	15.1
EN-TR	10.3	10.0	10.9	**11.3**
ZH-EN	13.5	13.7	**14.1**	13.7
EN-ZH	24.2	24.4	**25.6**	25.4
FR-DE	18.6	18.4	**19.9**	19.3
DE-FR	21.2	22.1	21.7	**22.6**

Table 4: Results for all task pairs in the WMT 2019 news-translation shared task where parallel training data is available.

ZERO-SHOT PIVOT. In the second, we use parallel evaluation datasets from the WMT shared tasks (consisting of (SRC, REF) pairs), and translate SRC → PIVOT → $\widehat{\text{TRG}}$, then measure performance on the resulting ($\widehat{\text{TRG}}$, REF) pairs (see below for more details), where the pivot and target language pair is a zero-shot translation task – this method is labeled ZERO-SHOT PARALLEL PIVOT.

Table 2 lists the WMT evaluation dataset that we use for each language pair. In the ZERO-SHOT PIVOT setting, the reference side of the dataset is used as input.

Table 3 shows global results for all parallel tasks and all zero-shot tasks, by system. Global scores are obtained by concatenating the segmented outputs for each translation direction, and computing the BLEU score against the corresponding concatenated, segmented reference translations. The results in table 3 are thus *tokenized* BLEU scores.

3.2 Parallel Tasks

In the following results, we report BLEU scores on de-tokenized output, and compute scores using sacrebleu [5]. Therefore, we expect BLEU scores to be equivalent to those used in the WMT automatic evaluation.

We note that across all but the lowest-resource tasks, the model with a unique decoder for each language outperforms all others. However, for EN→GU and EN→KK, the lowest-resource translation directions, the unique decoder model fails completely, probably because the unique parameters for KK and GU were not updated by a sufficient number of mini-batches (approximately 15,600 for EN→GU and 14,800 for EN→KK).

3.3 Zero-shot Translation Tasks

In order to test our models in the zero-shot setting, we adapt an evaluation technique that has recently been used for unsupervised MT – we translate from the source language into a pivot language, then back into the source language, and evaluate the score of the resulting source-language hypotheses against the original source (Lample

---

[5] BLEU+case.mixed+
lang.<src-lang>-<trg-lang>+
numrefs.1+smooth.exp+tok.<trg-lang>+
version.1.2.19

	PREPEND	EMB	DEC	ATTN
RU→CS→RU	20.9	**23.8**	20.8	21.0
RU→DE→RU	14.6	11.9	**16.4**	15.6
RU→EN→RU*	21.7	22.2	**23.9**	23.2
RU→FI→RU	11.2	**17.0**	12.1	11.6
RU→FR→RU	13.8	**15.4**	14.1	15.1
RU→GU→RU	**10.3**	9.6	3.5	5.1
RU→KK→RU	~~5.8~~	~~19.6~~	~~1.0~~	~~2.2~~
RU→LT→RU	16.9	**22.0**	16.5	16.6
RU→TR→RU	7.9	**10.2**	7.4	7.7
RU→ZH→RU	8.8	**10.5**	9.1	8.5

Table 5: Zero-shot translation results for RU→*→RU Note that BLEU scores are computed by translating SRC → PIVOT → $\widehat{SRC}$, and computing the score between SRC and $\widehat{SRC}$. Systems which do not pass the language identification filter are struck-through and removed from global evaluation. Note that parallel training data was available for RU→EN.

	PREPEND	EMB	DEC	ATTN
# Failed Pivot Tasks	3	31	1	1

Table 6: Out of 110 pivot translation tasks, how many failed the language identification check?

et al., 2018a). This technique allows us to evaluate for all possible translation directions in our multi-directional model.

Aware of the risk that the model simply copies through the original source segment instead of translating, we assert that at least 95% of pivot translations' language code is correctly detected by langid[6], and pairs which do not meet this criteria for any system are removed from the evaluation for all systems (not just for the system that failed). For all models except EMB only RU→KK→RU FI→LT→FI, and ZH→GU→ZH failed this test, but for the EMB model 31 of the 110 translation directions failed (see tables 6 and 7[7]. This result indicates that models which use language embeddings may have a more "fuzzy" representation of the output task, and are much more prone to copying than other approaches to multi-lingual MT. However, even for the languages which passed the language identification filter, we suspect that some copying is occurring

---

[6]https://github.com/saffsd/langid.py

[7]We conduct round trip translation on all 110 directions, but we only use directions that are (1) not available in the parallel training data, and (2) pass the language identification test to compute the global zero-shot translation performance

for the EMB system, because of the mismatch in results between the ZERO-SHOT PIVOT task and the SUPERVISED, ZERO-SHOT TED, and ZERO-SHOT PARALLEL PIVOT tasks (see table 3). Table 7 (in appendix) contains the results for all possible translation directions and all models in the ZERO-SHOT PIVOT evaluation setting.

3.3.1 Zero-Shot Evaluation on TED Talks Corpus

We conduct an additional evaluation on some of the language pairs from the TED Talks multi-parallel corpus (Ye et al., 2018), which has recently been used for the training and evaluation of multi-lingual models. We filter the dev and test sets of this corpus to find segments which have translations for all of EN, FR, RU, TR, DE, CS, LT, FI, and are at least 20 characters long, resulting in 606 segments. Because this corpus is preprocessed, we first de-tokenize and de-escape punctuation using sacremoses[8]. We then evaluate zero-shot translation for all possible pairs which do not occur in our parallel training data, aggregate results are shown in the third row of table 3.

3.4 Discussion

Our results show that a models with either (1) a completely unique decoders for each target language or (2) unique decoder attention parameters for each target language clearly outperform models with fully shared decoder parameters in our setting.

It is plausible that the language-independence of encoder output could be correlated with the amount of sharing in the decoder module. Because most non-English target tasks only have parallel training data in English, a unique decoder for those tasks only needs to learn to decode from English, not from every possible source task. However, our results show that the ATTN model, which partially shares parameters across target languages only slightly outperforms the DEC model globally, because of the improved performance of the ATTN model on the lowest-resource tasks (Table 4, Table 7 (in appendix)).

4   Related Work

Dong et al. (2015); Firat et al. (2016); Ha et al. (2016); Johnson et al. (2016) and others have

---

[8]https://github.com/alvations/sacremoses

shown that multi-way NMT systems can be created with minimal modification to the approach used for single-language-pair systems. Johnson et al. (2016) showed that simply prepending a target-task token to source inputs is enough to enable zero-shot translation between language pairs for which no parallel training data is available.

Our work is most similar to Sachan and Neubig (2018), where many different strategies for sharing decoder parameters are investigated for one-to-many translation models. However, their evaluation setting is constrained to one-to-many models which translate from English into two target languages, whereas our setting is more ambitious, performing multi-way translation between 11 languages. Blackwood et al. (2018) showed that using separate attention parameters for each task can improve the performance of multi-task MT models – this work was the inspiration for the ATTN setting in our experiments.

Several recent papers focus specifically upon improving the zero-shot performance of multi-lingual MT models (Chen et al., 2017; Arivazhagan et al., 2019; Gu et al., 2019; Lu et al., 2018; Al-Shedivat and Parikh, 2019; Sestorain et al., 2019).

Concurrently with this work, (Aharoni et al., 2019) evaluated a multiway MT system on a large number of language pairs using the TED talks corpus. However, they focus upon EN-* and *-EN, and do not test different model variants.

5 Conclusions and Future Work

We have presented results which are consistent with recent smaller-scale evaluations of multi-lingual MT systems, showing that assigning unique attention parameters to each target language in a multi-lingual NMT system is optimal when evaluating such a system globally. However, when evaluated on the individual task level, models which have unique decoder parameters for every target task tend to outperform other configurations, except when the amount of available training data is extremely small. We have also introduced two methods of evaluating zero-shot translation performance when parallel data is not available, and we conducted a large-scale evaluation of translation performance across all possible translation directions in the constrained setting of the WMT19 news-translation task.

In future work, we hope to continue studying how multi-lingual translation systems scale to realistic volumes of training data and large numbers of source and target tasks.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *Proceedings of NAACL*.

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

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3112–3122, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. 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, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: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. 2016. 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, San Diego, California. Association for Computational Linguistics.

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

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. Technical report, Google.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Belgium, Brussels. Association for Computational Linguistics.

Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *International Conference on Learning Representations*.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271, Belgium, Brussels. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Lierni Sestorain, Massimiliano Ciaramita, Christian Buck, and Thomas Hofmann. 2019. Zero-shot dual machine translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Qi Ye, Sachan Devendra, Felix Matthieu, Padmanabhan Sarguna, and Neubig Graham. 2018. When and why are pre-trained word embeddings useful for neural machine translation. In *HLT-NAACL*.

	CS	DE	EN	FI	FR	GU	KK	LT	RU	TR	ZH
PREPEND		15.7	19.6	11.4	11.1	8.7	3.6	16.6	17.4	7.8	7.5
EMB	**CS**	9.3	19.9	~~24.4~~	12.4	9.3	3.3	~~28.4~~	15.8	~~10.2~~	7.5
DEC		17.7	21.3	11.5	13.3	3.3	0.7	14.0	17.8	6.7	7.5
ATTN		17.5	21.6	11.6	13.8	4.5	1.8	14.4	17.4	7.2	7.6
PREPEND	22.3		27.1	16.4	25.3	11.2	5.3	18.3	18.1	12.4	12.3
EMB	~~41.5~~	**DE**	27.6	~~38.0~~	25.1	~~19.4~~	8.5	~~40.7~~	23.6	~~29.9~~	~~19.6~~
DEC	23.4		29.8	15.6	25.8	4.0	0.9	16.2	19.1	11.2	12.2
ATTN	22.8		29.0	15.9	26.3	6.4	2.7	17.1	18.0	11.2	12.1
PREPEND	35.4	37.1		24.6	34.5	22.1	9.3	29.2	32.9	23.3	25.6
EMB	36.5	37.4	**EN**	25.9	~~35.0~~	21.5	9.3	30.4	33.9	24.2	26.6
DEC	35.8	37.5		25.8	32.6	10.1	1.6	29.7	33.2	22.8	26.3
ATTN	36.9	36.6		25.9	34.4	15.7	6.2	30.3	33.9	23.6	26.8
PREPEND	12.1	11.0	14.3		7.2	5.6	2.6	~~13.1~~	9.2	6.6	6.2
EMB	~~19.9~~	~~7.9~~	14.7	**FI**	8.2	6.2	4.2	~~23.8~~	11.1	~~12.0~~	~~6.8~~
DEC	11.2	11.7	15.4		9.8	3.0	0.5	10.7	9.9	6.2	5.8
ATTN	12.2	11.5	15.0		10.0	4.2	1.7	10.9	9.8	6.5	5.8
PREPEND	25.6	32.7	31.9	17.8		17.1	7.8	20.5	22.9	16.0	15.3
EMB	26.0	32.0	~~36.5~~	20.4	**FR**	12.4	5.0	~~24.5~~	22.5	15.6	14.1
DEC	25.7	32.9	33.9	18.4		5.9	1.5	20.2	23.6	14.4	15.1
ATTN	26.0	33.2	34.3	19.5		8.5	4.5	21.0	24.6	15.5	15.4
PREPEND	5.1	5.7	8.2	4.2	4.0		2.0	5.3	5.2	3.5	3.9
EMB	~~5.3~~	4.6	7.6	5.4	4.4	**GU**	1.6	~~6.2~~	3.6	4.2	3.7
DEC	3.5	3.5	5.5	3.2	3.2		0.5	3.6	3.4	2.9	2.7
ATTN	4.9	5.1	7.6	4.4	4.5		1.0	4.6	4.5	4.0	4.0
PREPEND	4.9	4.8	7.1	3.4	2.8	4.3		4.6	4.9	4.4	3.9
EMB	3.4	3.4	6.4	3.8	3.5	~~1.9~~	**KK**	3.9	~~3.6~~	4.5	2.4
DEC	1.7	1.8	2.2	1.6	1.5	1.1		1.5	1.7	1.7	1.2
ATTN	3.9	4.0	5.3	3.2	3.2	2.3		3.6	3.9	4.3	3.2
PREPEND	18.8	14.5	17.9	13.8	10.4	9.9	4.7		16.8	10.0	9.0
EMB	~~30.5~~	~~12.0~~	18.3	~~30.7~~	10.8	10.6	~~6.8~~	**LT**	18.2	~~19.7~~	13.7
DEC	16.7	13.6	18.9	12.8	11.6	3.7	0.9		16.1	8.0	8.5
ATTN	17.0	13.9	18.8	12.5	12.4	5.5	1.9		15.7	8.7	9.2
PREPEND	20.9	14.6	21.7	11.2	13.8	10.3	~~5.8~~	16.9		7.9	8.8
EMB	23.8	11.9	22.2	17.0	15.4	9.6	~~19.6~~	~~22.0~~	**RU**	10.2	10.5
DEC	20.8	16.4	23.9	12.1	14.1	3.5	~~1.0~~	16.5		7.5	9.1
ATTN	21.0	15.6	23.2	11.6	15.2	5.1	~~2.2~~	16.6		7.7	8.5
PREPEND	9.1	8.2	13.3	7.4	7.4	8.8	5.6	8.9	7.1		6.4
EMB	~~12.7~~	7.2	12.6	~~14.6~~	7.5	6.7	4.1	~~17.3~~	6.6	**TR**	6.4
DEC	7.2	7.6	13.1	6.5	6.8	2.5	0.7	6.4	5.6		5.3
ATTN	7.3	8.1	13.4	6.6	7.3	3.9	1.9	6.7	5.6		5.3
PREPEND	20.4	19.6	29.0	17.1	17.4	~~18.2~~	8.4	20.2	19.5	17.4	
EMB	20.1	16.9	29.4	19.6	17.8	11.9	6.6	22.8	18.3	16.7	**ZH**
DEC	19.2	19.4	30.2	16.6	17.6	7.2	2.2	19.5	20.1	16.3	
ATTN	19.8	20.4	30.0	16.7	18.2	11.0	5.0	18.6	19.4	17.1	

Table 7: Pivot-based translation results in all directions, for all models. Rows indicate source language, columns indicate pivot language. For example, cell $(1,2)$ contains the results for CS→DE→CS. Runs which did not pass the language identification filter are struck-through. The MT-matrix (http://matrix.statmt.org/matrix) was the inspiration for this rendering.