# Multi-task Learning for Multilingual Neural Machine Translation

**Yiren Wang**[†,*], **ChengXiang Zhai**[†], **Hany Hassan Awadalla**[‡]
[†]University of Illinois at Urbana-Champaign
[‡]Microsoft
[†]{yiren, czhai}@illinois.edu
[‡]hanyh@microsoft.com



## Abstract

While monolingual data has been shown to be useful in improving bilingual neural machine translation (NMT), effectively and efficiently leveraging monolingual data for Multilingual NMT (MNMT) systems is a less explored area. In this work, we propose a multi-task learning (MTL) framework that jointly trains the model with the translation task on bitext data and two denoising tasks on the monolingual data. We conduct extensive empirical studies on MNMT systems with 10 language pairs from WMT datasets. We show that the proposed approach can effectively improve the translation quality for both high-resource and low-resource languages with large margin, achieving significantly better results than the individual bilingual models. We also demonstrate the efficacy of the proposed approach in the zero-shot setup for language pairs without bitext training data. Furthermore, we show the effectiveness of MTL over pre-training approaches for both NMT and cross-lingual transfer learning NLU tasks; the proposed approach outperforms massive scale models trained on single task.

## 1 Introduction

Multilingual Neural Machine Translation (MNMT), which leverages a single NMT model to handle the translation of multiple languages, has drawn research attention in recent years (Dong et al., 2015; Firat et al., 2016a; Ha et al., 2016; Johnson et al., 2017; Arivazhagan et al., 2019). MNMT is appealing since it greatly reduces the cost of training and serving separate models for different language pairs (Johnson et al., 2017). It has shown great potential in knowledge transfer among languages, improving the translation quality for low-resource and zero-shot language pairs (Zoph et al., 2016; Firat et al., 2016b; Arivazhagan et al., 2019).

[*]Work done while interning at Microsoft.

Previous works on MNMT has mostly focused on model architecture design with different strategies of parameter sharing (Firat et al., 2016a; Blackwood et al., 2018; Sen et al., 2019) or representation sharing (Gu et al., 2018). Existing MNMT systems mainly rely on bitext training data, which is limited and costly to collect. Therefore, effective utilization of monolingual data for different languages is an important research question yet is less studied for MNMT.

Utilizing monolingual data (more generally, the unlabeled data) has been widely explored in various NMT and natural language processing (NLP) applications. Back translation (BT) (Sennrich et al., 2016), which leverages a target-to-source model to translate the target-side monolingual data into source language and generate pseudo bitext, has been one of the most effective approaches in NMT. However, well trained NMT models are required to generate back translations for each language pair, it is computationally expensive to scale in the multilingual setup. Moreover, it is less applicable to low-resource language pairs without adequate bitext data. Self-supervised pre-training approaches (Radford et al., 2018; Devlin et al., 2019; Conneau and Lample, 2019; Lewis et al., 2019; Liu et al., 2020), which train the model with denoising learning objectives on the large-scale monolingual data, have achieved remarkable performances in many NLP applications. However, catastrophic forgetting effect (Thompson et al., 2019), where finetuning on a task leads to degradation on the main task, limits the success of continuing training NMT on models pre-trained with monolingual data. Furthermore, the separated pre-training and finetuning stages make the framework less flexible to introducing additional monolingual data or new languages into the MNMT system.

In this paper, we propose a multi-task learning (MTL) framework to effectively utilize monolin-

gual data for MNMT. Specifically, the model is jointly trained with translation task on multilingual parallel data and two auxiliary tasks: masked language modeling (MLM) and denoising auto-encoding (DAE) on the source-side and target-side monolingual data respectively. We further present two simple yet effective scheduling strategies for the multilingual and multi-task framework. In particular, we introduce a dynamic temperature-based sampling strategy for the multilingual data. To encourage the model to keep learning from the large-scale monolingual data, we adopt dynamic noising ratio for the denoising objectives to gradually increase the difficulty level of the tasks.

We evaluate the proposed approach on a large-scale multilingual setup with 10 language pairs from the WMT datasets. We study three English-centric multilingual systems, including many-to-English, English-to-many, and many-to-many. We show that the proposed MTL approach significantly boosts the translation quality for both high-resource and low-resource languages. Furthermore, we demonstrate that MTL can effectively improve the translation quality on zero-shot language pairs with no bitext training data. In particular, MTL achieves even better performance than the pivoting approach for multiple low-resource language pairs. We further show that MTL outperforms pre-training approaches on both NMT tasks as well as cross-lingual transfer learning for NLU tasks, despite being trained on very small amount of data in comparison to pre-training approaches.

The contributions of this paper are as follows. First, we propose a new MTL approach to effectively utilize monolingual data for MNMT. Second, we introduce two simple yet effective scheduling strategies, namely the dynamic temperature-based sampling and dynamic noising ratio strategy. Third, we present detailed ablation studies to analyze various aspects of the proposed approach. Finally, we demonstrate for the first time that MNMT with MTL models can be effectively used for cross-lingual transfer learning for NLU tasks with similar or better performance than the state-of-the-art massive scale pre-trained models using single task.

## 2 Background

**Neural Machine Translation** NMT adopts the sequence-to-sequence framework, which consists of an encoder and a decoder network built upon deep neural networks (Sutskever et al., 2014; Bahdanau et al., 2014; Gehring et al., 2017; Vaswani et al., 2017). The input source sentence is mapped into context representations in a continuous representation space by the encoder, which are then fed into the decoder to generate the output sentence. Given a language pair $(x, y)$, the objective of the NMT model training is to maximize the conditional probability $P(y|x;\theta)$ of the target sentence given the source sentence.

NMT heavily relies on high-quality and large-scale bitext data. Various strategies have been proposed to augment the limited bitext by leveraging the monolingual data. Back translation (Sennrich et al., 2016) utilizes the target-side monolingual data. Self learning (Zhang and Zong, 2016) leverages the source-side monolingual data. Dual learning paradigms utilize monolingual data in both source and target language (He et al., 2016; Wang et al., 2019; Wu et al., 2019). While these approaches can effectively improve the NMT performance, they have two limitations. First, they introduce additional cost in model training and translation generation, and therefore are less efficient when scaling to the multilingual setting. Second, back translation requires a good baseline model with adequate bitext data to start from, which limits its efficiency on low-resource settings.

**Multilingual NMT** MNMT aims to train a single translation model that translates between multiple language pairs (Firat et al., 2016a; Johnson et al., 2017). Previous works explored the model architecture design with different parameter sharing strategies, such as partial sharing with shared encoder (Dong et al., 2015; Sen et al., 2019), shared attention (Firat et al., 2016a), task-specific attention (Blackwood et al., 2018), and full model sharing with language identifier (Johnson et al., 2017; Ha et al., 2016; Arivazhagan et al., 2019). There are also extensive studies on representation sharing that shares lexical, syntactic, or sentence level representations across different languages (Zoph et al., 2016; Nguyen and Chiang, 2017; Gu et al., 2018). The models in these works rely on bitext for training, and the largely available monolingual data has not been effectively leveraged.

**Self-supervised Learning** This work is motivated by the recent success of self-supervised learning for NLP applications (Radford et al., 2018; Devlin et al., 2019; Lample et al., 2018a,b; Conneau and Lample, 2019; Lewis et al., 2019; Liu et al.,

2020). Different denoising objectives have been designed to train the neural networks on large-scale unlabeled text. In contrast to previous work in pre-training with separated self-supervised pre-training and supervised finetuning stages, we focus on a multi-task setting to *jointly* train the MNMT model on both bitext and monolingual data.

**Multi-task Learning** Multi-task learning (MTL) (Caruana, 1997), which trains the model on several related tasks to encourage representation sharing and improve generalization performance, has been successfully used in many different machine learning applications (Collobert and Weston, 2008; Deng et al., 2013; Ruder, 2017). In the context of NMT, MTL has been explored mainly to inject linguistic knowledge (Luong et al., 2015; Niehues and Cho, 2017; Eriguchi et al., 2017; Zaremoodi and Haffari, 2018; Kiperwasser and Ballesteros, 2018) with tasks such as part-of-speech tagging, dependency parsing, semantic parsing, etc. In this work, we instead focus on auxiliary self-supervised learning tasks to leverage the monolingual data.

## 3 Approach

### 3.1 Multi-task Learning

The main task in the MTL framework is the translation task trained on bitext corpora $D_B$ of sentence pairs $(x, y)$ with the cross-entropy loss:

$$\mathcal{L}_{MT} = \mathbb{E}_{(x,y)\sim D_B}[-\log P(y|x)] \quad (1)$$

With the large amount of monolingual data in different languages, we can train language models on both source-side [1] and target-side languages. We introduce two denoising language modeling tasks to help improve the quality of the translation model: the masked language model (MLM) task and the denoising auto-encoding (DAE) task.

**Masked Language Model** In the masked language model (MLM) task (Devlin et al., 2019), sentences with tokens randomly masked are fed into the model and the model attempts to predict the masked tokens based on their context. MLM is beneficial for learning deep bidirectional representations. We introduce MLM as an auxiliary task to improve the quality of the encoder representations especially for the low-resource languages. As is illustrated in Figure 1(a), we add an additional output layer to the encoder of the translation model and train the encoder with MLM on source-side monolingual data. The output layer is dropped during inference. The cross entropy loss for predicting the masked tokens is denoted as $\mathcal{L}_{MLM}$.

Following BERT (Devlin et al., 2019), we randomly sample $R_M\%$ units in the input sentences and replace them with a special `[MASK]` token. A unit can either be a subword token, or a word consists of one or multiple subword tokens. We refer to them as token-level and word-level MLM.

**Denoising Auto-Encoding (DAE)** Denoising auto-encoding (DAE) (Vincent et al., 2008) has been demonstrated to be an effective strategy for unsupervised NMT (Lample et al., 2018a,b). Given a monolingual corpus $D_M$ and a stochastic noising model $C$, DAE minimizes the reconstruction loss as shown in Eqn 2:

$$\mathcal{L}_{DAE} = \mathbb{E}_{x\sim D_M}[-\log P(x|C(x))] \quad (2)$$

As is illustrated in Figure 1(b), we train all model parameters with DAE on the target-side monolingual data. Specifically, we feed the target-side sentence to the noising model $C$ and append the corresponding language ID symbol; the model then attempts to reconstruct the original sentence.

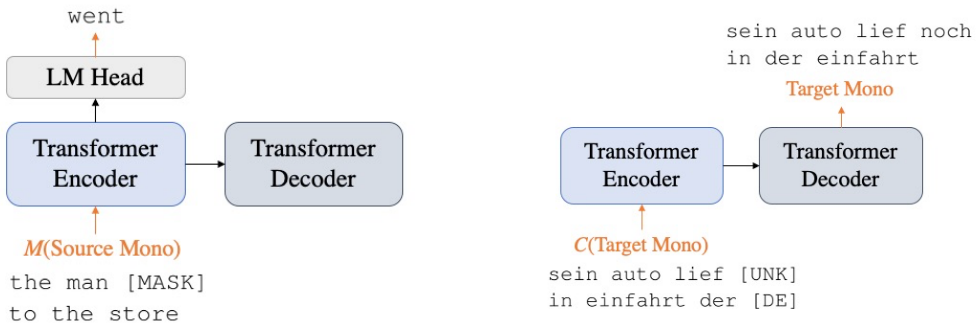
We introduce three types of noises for the noising model $C$. 1) *Text Infilling* (Lewis et al., 2019): Following (Liu et al., 2020), we randomly sample $R_D\%$ text spans with span lengths drawn from a Poisson distribution ($\lambda = 3.5$). We replace all words in each span with a single blanking token. 2) *Word Drop & Word Blank*: we randomly sample words from each input sentence, which are either removed or replaced with blanking tokens for each token position. 3) *Word Swapping*: we slightly shuffle the order of words in the input sentence. Following (Lample et al., 2018a), we apply a random permutation $\sigma$ with condition $|\sigma(i) - i| \leq k, \forall i \in \{1, n\}$, where $n$ is the length of the input sentence, and $k = 3$ is the maximum swapping distance.

**Joint Training** In the training process, the two self-learning objectives are combined with the cross-entropy loss for the translation task:

$$\mathcal{L} = \mathcal{L}_{MT} + \mathcal{L}_{MLM} + \mathcal{L}_{DAE} \quad (3)$$

In particular, we use bitext data for the translation objective, source-side monolingual data for MLM,

[1] For the English-to-Many translation model, the source-side language is English; for Many-to-English and Many-to-Many, it refers the set of all other languages. Similarly for the target-side language.




(a) Masked Language Model (MLM)

(b) Denoising Auto-Encoding (DAE)

Figure 1: Illustration of the auxiliary tasks with monolingual data

and target-side monolingual data for the DAE objective. A language ID symbol `[LID]` of the target language is appended to the input sentence in the translation and DAE tasks.

### 3.2 Task Scheduling

The scheduling of tasks and data associated with the task is important for multi-task learning. We further introduce two simple yet effective scheduling strategies in the MTL framework.

**Dynamic Data Sampling** One serious yet common problem for MNMT is data imbalance across different languages. Training the model with the true data distribution would starve the low-resource language pairs. Temperature-based batch balancing (Arivazhagan et al., 2019) is demonstrated to be an effective heuristic to ease the problem. For language pair $l$ with bitext corpus $D_l$, we sample instances with probability proportional to $(\frac{|D_l|}{\sum_k |D_k|})^{\frac{1}{T}}$, where $T$ is the sampling temperature.

While MNMT greatly improves translation quality for low-resource languages, performance deterioration is generally observed for high resource languages. One hypothesized reason is that the model might converge before well trained on high-resource data (Bapna and Firat, 2019). To alleviate this problem, we introduce a simple heuristic to feed more high-resource language pairs in the early stage of training and gradually shift more attention to the low-resource languages. To achieve this, we modify the sampling strategy by introducing dynamic sampling temperature $T(k)$ as a function of the number of training epochs $k$. We use a simple linear functional form for $T(k)$:

$$T(k) = \min\left(T_m, (k-1)\frac{T_m - T_0}{N} + T_0\right) \quad (4)$$

Where $T_0$ and $T_m$ are the initial and maximum value for sampling temperature respectively. $N$ is the number of warm-up epochs. The sampling temperature starts from a smaller value $T_0$, resulting in sampling leaning towards true data distribution. $T(k)$ gradually increases in the training process to encourage over-sampling low-resource languages more to avoid them getting starved.

**Dynamic Noising Ratio** We further schedule the difficulty level of MLM and DAE from easier to more difficult. The main motivation is that training algorithms perform better when starting with easier tasks and gradually move to harder ones as promoted in curriculum learning (Elman, 1993). Furthermore, increasing the learning difficulty can potentially help avoid saturation and encourage the model to keep learning from abundant data.

Given the monolingual data, the difficulty level of MLM and DAE tasks mainly depends on the noising ratio. Therefore, we introduce dynamic noising ratio $R(k)$ as a function of training steps:

$$R(k) = \min\left(R_m, (k-1)\frac{R_m - R_0}{M} + R_0\right) \quad (5)$$

Where $R_0$ and $R_m$ are the lower and upper bound for noising ratio respectively and $M$ is the number of warm-up epochs. Noising ratio $R$ refers to the masking ratio $R_M$ in MLM and the blanking ratio $R_D$ of the *Text Infilling* task for DAE.

## 4 Experimental Setup

### 4.1 Data

We evaluate MTL on a multilingual setting with 10 languages to and from English (En), including French (Fr), Czech (Cs), German (De), Finnish

(Fi), Latvian (Lv), Estonian (Et), Romanian (Ro), Hindi (Hi), Turkish (Tr) and Gujarati (Gu).

**Bitext Data** The bitext training data comes from the WMT corpus. Detailed desciption and statistics can be found in Appendix A.

**Monolingual Data** The monolingual data we use is mainly from NewsCrawl[2]. We apply a series of filtration rules to remove the low-quality sentences, including duplicated sentences, sentences with too many punctuation marks or invalid characters, sentences with too many or too few words, etc. We randomly select 5M filtered sentences for each language. For low-resource languages without enough sentences from NewsCrawl, we leverage data from CCNet (Wenzek et al., 2019).

**Back Translation** We use the target-to-source bilingual models to back translate the target-side monolingual sentences into the source domain for each language pair. The synthetic parallel data from back translation is mixed and shuffled with bitext and used together for the translation objective in training. We use the same monolingual data for back translation as the multi-task learning in all our experiments for fair comparison.

### 4.2 Model Configuration

We use Transformer for all our experiments using the PyTorch implementation[3] (Ott et al., 2019). We adopt the `transformer_big` setting (Vaswani et al., 2017) with a 6-layer encoder and decoder. The dimensions of word embeddings, hidden states, and non-linear layer are set as 1024, 1024 and 4096 respectively, the number of heads for multi-head attention is set as 16. We use a smaller model setting for the bilingual models on low-resource languages Tr, Hi and Gu (with 3 encoder and decoder layers, 256 embedding and hidden dimension) to avoid overfitting and acquire better performance.

We study three multilingual translation scenarios including many-to-English (X→En), English-to-many (En→X) and many-to-many (X→X). For the multilingual model, we adopt the same Transformer architecture as the bilingual setting, with parameters fully shared across different language pairs. A target language ID token is appended to each input sentence.

[2] http://data.statmt.org/news-crawl/
[3] https://github.com/pytorch/fairseq

### 4.3 Training and Evaluation

All models are optimized with Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$. We set the learning rate schedule following (Vaswani et al., 2017) with initial learning rate $5 \times 10^{-4}$. Label smoothing (Szegedy et al., 2016) is adopted with 0.1. The models are trained on 8 V100 GPUs with a batch size of 4096 and the parameters are updated every 16 batches. During inference, we use beam search with a beam size of 5 and length penalty 1.0. The BLEU score is measured by the de-tokenized case-sensitive SacreBLEU[4] (Post, 2018).

## 5 Results

### 5.1 Main Results

We compare the performance of the bilingual models (*Bilingual*), multilingual models trained on bitext only, trained on both bitext and back translation (*+BT*) and trained with the proposed multi-task learning (*+MTL*). Translation results of the 10 languages translated to and from English are presented in Table 1 and 2 respectively. We can see that:

1. *Bilingual vs. Multilingual*: The multilingual baselines perform better on lower-resource languages, but perform worse than individual bilingual models on high-resource languages like Fr, Cs and De. This is in concordance with the previous observations (Arivazhagan et al., 2019) and is consistent across the three multilingual systems (i.e., X→En, En→X and X→X).
2. *Multi-task learning*: Models trained with multi-task learning (+MTL) significantly outperform the multilingual baselines for all the languages pairs in all three multilingual systems, demonstrating the effectiveness of the proposed framework.
3. *Back Translation*: With the same monolingual corpus, MTL achieves better performance on some language pairs (e.g. Fr→En, Gu→En), while getting outperformed on some others, especially on the En→X direction. However, back translation is computationally expensive as it involves the additional procedure of training 10 bilingual models (20 for the X→X system) and generating translations for each monolingual sentence. Combining MTL with BT (+BT+MTL) introduces further improvements for most language pairs without using any additional monolingual data. This suggests

[4] SacreBLEU signatures: BLEU+case.mixed+lang.$l1-$l2 numrefs.1+smooth.exp+test.$SET+tok.13a+version.1.4.3, where $l1, $l2 are the language code (Table 9), $SET is the corresponding test set for the language pair.

| Test Set | Fr wmt15 | Cs wmt18 | De wmt18 | Fi wmt18 | Lv wmt17 | Et wmt18 | Ro wmt16 | Hi wmt14 | Tr wmt18 | Gu wmt19 |
|---|---|---|---|---|---|---|---|---|---|---|
| Bilingual | 36.2 | 28.5 | 40.2 | 19.2 | 17.5 | 19.7 | 29.8 | 14.1 | 15.1 | 9.3 |
| X → En | 34.6 | 28.0 | 39.7 | 20.1 | 19.6 | 23.9 | 33.2 | 20.5 | 21.3 | 16.1 |
| + MTL | 36.4 | 31.5 | 42.3 | 23.0 | 22.1 | 28.7 | 37.9 | 24.8 | 25.7 | 22.3 |
| + BT | 35.3 | 31.2 | 44.3 | 23.4 | 21.4 | 29.2 | 37.9 | 27.2 | 25.5 | 21.5 |
| + BT + MTL | 35.3 | 31.9 | 45.4 | 23.8 | 22.4 | 30.5 | 39.1 | 28.7 | 27.6 | 23.5 |
| X → X | 33.9 | 28.1 | 39.0 | 19.9 | 19.5 | 24.5 | 33.7 | 22.4 | 22.0 | 17.2 |
| + MTL | 35.1 | 29.6 | 40.1 | 21.7 | 21.3 | 27.3 | 36.8 | 23.9 | 25.2 | 23.3 |
| + BT | 34.3 | 30.6 | 43.7 | 22.8 | 20.9 | 28.0 | 37.3 | 26.4 | 25.5 | 22.5 |
| + BT + MTL | 35.3 | 31.2 | 43.7 | 23.1 | 21.5 | 29.5 | 38.1 | 27.5 | 26.2 | 23.4 |

Table 1: BLEU scores of 10 languages → English translation with bilingual, X→En and X→X systems. The languages are arranged from high-resource (left) to low-resource (right).

| Test Set | Fr wmt15 | Cs wmt18 | De wmt18 | Fi wmt18 | Lv wmt17 | Et wmt18 | Ro wmt16 | Hi wmt14 | Tr wmt18 | Gu wmt19 |
|---|---|---|---|---|---|---|---|---|---|---|
| Bilingual | 36.3 | 22.3 | 40.2 | 15.2 | 16.5 | 15.0 | 23.0 | 12.2 | 13.3 | 7.9 |
| En → X | 33.5 | 20.8 | 39.0 | 14.9 | 18.0 | 19.8 | 25.5 | 12.4 | 15.7 | 11.9 |
| + MTL | 33.8 | 21.7 | 39.8 | 15.2 | 18.5 | 21.1 | 26.5 | 16.1 | 17.6 | 15.4 |
| + BT | 35.9 | 22.5 | 41.5 | 17.3 | 21.8 | 23.0 | 28.8 | 19.1 | 18.6 | 15.5 |
| + BT + MTL | 36.1 | 23.6 | 42.0 | 17.7 | 22.4 | 24.0 | 29.8 | 19.8 | 19.4 | 17.8 |
| X → X | 32.2 | 19.4 | 37.3 | 14.5 | 17.5 | 19.6 | 25.4 | 13.9 | 16.3 | 12.0 |
| + MTL | 33.3 | 20.9 | 39.2 | 15.6 | 19.3 | 21.1 | 26.8 | 16.5 | 18.1 | 15.5 |
| + BT | 35.9 | 22.0 | 40.0 | 16.3 | 21.1 | 22.8 | 28.7 | 19.0 | 18.2 | 15.9 |
| + BT + MTL | 35.8 | 22.4 | 41.2 | 16.9 | 21.7 | 23.2 | 29.7 | 19.2 | 18.7 | 16.0 |

Table 2: BLEU scores of English → 10 languages translation with bilingual, En→X and X→X systems. The languages are arranged from high-resource (left) to low-resource (right).

that when there is enough computation budget for BT, MTL can still be leveraged to provide good complementary improvement.

### 5.2 Zero-shot Translation

We further evaluate the proposed approach on zero-shot translation of non English-centric language pairs. We compare the performances of the pivoting method, the X→X baseline system, X→X with BT, and with MTL. For the pivoting method, the source language is translated into English first, and then translated into the target language (De Gispert and Marino, 2006; Utiyama and Isahara, 2007). We evaluate on a group of high-resource languages with a multi-way parallel test set for De, Cs, Fr and En, constructed by newstest2009 with 3027 sentences and that of a group of low-resource languages Et, Hi, Tr and Hi (995 sentences). The results are shown in Table 3 and 4 respectively.

Utilizing monolingual data with MTL significantly improves the zero-shot translation quality of the X→X system, further demonstrating the effectiveness of the proposed approach. In particular, MTL achieves significantly better results than the pivoting approach on the high-resource pair Cs→De and almost all low-resource pairs. Furthermore, leveraging monolingual data through BT does not perform well for many low-resource language pairs, resulting in comparable and even downgraded performances. We conjecture that this is related to the quality of the back translations. MTL helps overcome such limitations with the auxiliary self-supervised learning tasks.

| | De→Fr | Fr→De | Cs→De | De→Cs |
|---|---|---|---|---|
| Pivoting | 22.1 | 19.1 | 17.5 | 15.9 |
| X→X | 15.1 | 11.9 | 15.5 | 15.2 |
| + BT | 19.7 | 7.4 | 17.0 | 7.8 |
| + BT + MTL | 20.1 | 12.2 | 19.7 | 12.0 |

Table 3: Zero-shot translation performances on high-resource language pairs.

### 5.3 MTL vs. Pre-training

We also compare MTL with mBART (Liu et al., 2020), the state-of-the-art multilingual pre-training method for NMT. We adopt the officially released mBART model pre-trained on CC25 corpus[5] and

[5] https://dl.fbaipublicfiles.com/fairseq/models/mbart/mbart.CC25.tar.gz

| | Et→Hi | Hi→Et | Hi→Tr | Tr→Hi |
|---|---|---|---|---|
| Pivoting | 8.1 | 7.1 | 3.9 | 6.0 |
| X→X | 5.0 | 5.4 | 2.5 | 3.6 |
| + BT | 4.9 | 5.5 | 4.5 | 5.7 |
| + BT + MTL | 9.3 | 8.1 | 5.7 | 9.4 |
| | Et→Tr | Tr→Et | Hi→Lv | Lv→Hi |
| Pivoting | 7.1 | 7.8 | 8.6 | 7.9 |
| X→X | 7.8 | 8.8 | 6.1 | 4.7 |
| + BT | 7.3 | 7.5 | 7.3 | 6.1 |
| + BT + MTL | 7.8 | 10.5 | 8.2 | 8.6 |

Table 4: Zero-shot translation performances on low-resource language pairs.

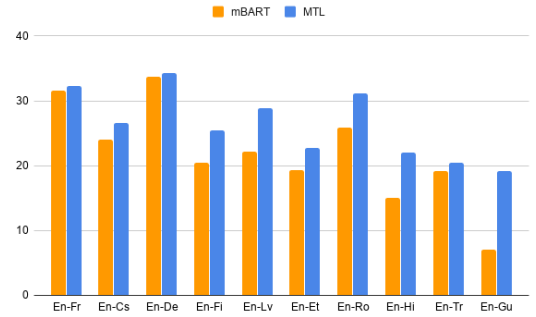



Figure 2: Comparison with mBART on En→X language pairs. BLEU scores are reported on the full individual validation set.

finetune the model on the same bitext training data used in MTL for each language pair. As shown in Figure 2, MTL outperforms mBART on all language pairs. This suggests that in the scenario of NMT, jointly training the model with MT task and self-supervised learning tasks could be a better task design than the separated pre-training and finetuning stages. It is worth noting that mBart is utilizing much more monolingual data; for example, it uses 55B English tokens and 10B French tokens, while our approach is using just 100M tokens each. This indicates that MTL is more data efficient.

### 5.4 Multi-task Objectives

We present ablation study on the learning objectives of the multi-task learning framework. We compare performance of multilingual baseline model with translation objective only, jointly learning translation with MLM, jointly learning translation with DAE, and the combination of all objectives. Table 5 shows the results on a high-resource pair De↔En and low-resource pair Tr↔En. We can see that introducing MLM or DAE can both effectively improve the performance of multilingual systems, and the combination of both yields the best performance. We also observe that MLM is more beneficial for '→En' compared with 'En→' direction, especially for the low-resource languages. This is in concordance with our intuition that the MLM objective contributes to improving the encoder quality and source-side language modeling for low-resource languages.

| Systems | De→En X→En | De→En X→X | Tr→En X→En | Tr→En X→X |
|---|---|---|---|---|
| Multilingual | 36.5 | 36.1 | 20.0 | 20.9 |
| + MLM | 36.3 | 36.6 | 21.0 | 21.4 |
| + DAE | 37.7 | 37.8 | 21.7 | 22.6 |
| + MLM + DAE | 38.7 | 37.6 | 22.9 | 23.7 |
| Systems | En→De En→X | En→De X→X | En→Tr En→X | En→Tr X→X |
| Multilingual | 33.0 | 32.0 | 16.4 | 17.0 |
| + MLM | 32.9 | 32.6 | 16.9 | 17.2 |
| + DAE | 33.7 | 33.7 | 17.3 | 18.2 |
| + MLM + DAE | 34.2 | 33.6 | 18.0 | 18.3 |

Table 5: Comparison of different multi-task learning objectives on De-En and Tr-En translation. BLEU scores are reported on the full individual validation set.

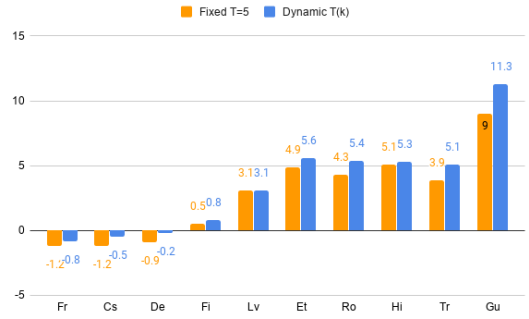



Figure 3: Performance gain of data sampling strategies on the X→En system. Results are reported as ΔBLEU relative to the corresponding bilingual baseline on validation sets. The languages are arranged from high-resource (left) to low-resource (right).

### 5.5 Dynamic Sampling Temperature

We study the effectiveness of the proposed dynamic sampling strategy. We compare multilingual systems using a fixed sampling temperature $T = 5$ with systems using dynamic temperature $T(k)$ defined in Equation 4. We set $T_0 = 1, T_m = 5, N = 5$, which corresponds to gradually increasing the temperature from 1 to 5 with 5 training epochs and saturate to $T = 5$ afterwards. The results for X→En and En→X systems are presented in Figure 3 and 4 respectively, where we report ΔBLEU relative to their corresponding bilingual baseline

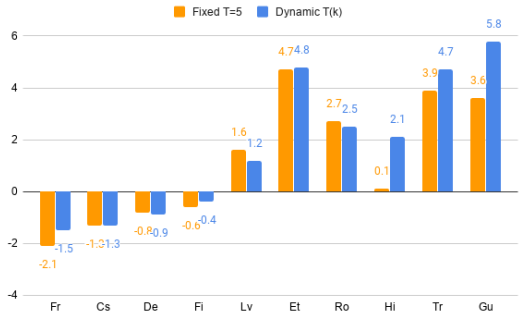



Figure 4: Performance of data sampling strategies on the En→X system.

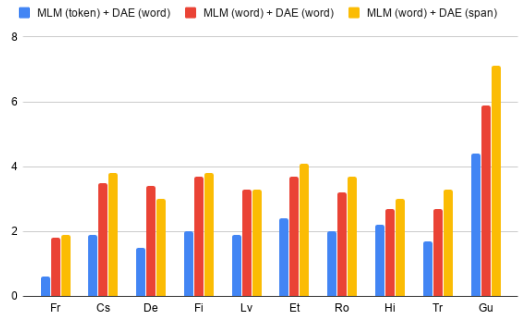



Figure 5: Performance of different noising schemes on the X→En system. Results are reported as ΔBLEU relative to the multilingual X→En baseline on validation sets. The languages are arranged from high-resource (left) to low-resource (right).

model that was evaluated on the individual validation sets for each language pairs. The dynamic temperature strategy improves the quality for high-resource language pairs (e.g. Fr→En, De→En, En→Fr), while introducing minimum effect for mid-resource languages (Lv). Surprisingly, the proposed strategy also greatly boosts performance for low-resource languages Tr and Gu, with over +1 BLEU gain for both to and from English direction.

### 5.6 Noising Scheme

We study the effect of different noising schemes in the MLM and DAE objectives. As introduced in Section 3.1, we have token-level and word-level masking scheme for MLM depending on the unit of masking. We also have two noising schemes for DAE, where the *Text Infilling* task blanks a span of words (span-level), and the *Word Blank* task blanks the input sentences at word-level. We compare performance of these different noising schemes on X→En system as shown in Figure 5.

We report ΔBLEU relative to the multilingual X→En baseline on the corresponding language pairs for each noising scheme. As we can see, the model benefits most from the word-level MLM and the span-level *Text Infilling* task for DAE. This is in concordance with the intuition that the *Text Infilling* task teaches the model to predict the length of masked span and the exact tokens at the same time, making it a harder task to learn. We use the word-level MLM and span-level DAE as the best recipe for our MTL framework.

| | De | Lv | Et | Hi | Tr |
|---|---|---|---|---|---|
| Bilingual | 32.9 | 23.4 | 18.7 | 12.9 | 16.1 |
| X→En + BT | 35.2 | 30.1 | 28.3 | 18.7 | 24.7 |
| + MTL | 36.9 | 31.9 | 31.4 | 21.6 | 27.4 |
| + Dynamic | 37.0 | 32.4 | 32.0 | 21.7 | 27.5 |

Table 6: BLEU scores of dynamic noising strategy on X→En translation system with large-scale monolingual data setting on validation sets.

### 5.7 Noising Ratio Scheduling

In our initial experiments, we found that the dynamic noising ratio strategy does not effectively improve the performance. We suspect that it is due to the limitation of data scale. We experiment with a larger scale setting by increasing the amount of monolingual data from 5M sentences for each language to 20M. For low-resource languages without enough data, we take the full available amount (18M for Lv, 11M for Et, 5.2M for Gu).

Table 6 shows results on X→En MNMT model with large-scale monolingual data setting. We compare the performance of multilingual with back translation baseline, a model with MTL and a model with both MTL and dynamic noising ratio. For the dynamic noising ratio, we set the masking ratio for MLM to increase from 10% to 20% and blanking ratio for DAE to increase from 20% to 40%. As we can see, the dynamic noising strategy helps boost performance for mid-resource languages like Lv and Et, while introducing no negative effect to other languages. For future study, we would like to cast the dynamic noising ratio over different subsets of monolingual datasets to prevent the model from learning to copy and memorize.

### 5.8 MTL for Cross-Lingual Transfer Learning for NLU

Large scale pre-trained cross-lingual language models such as mBERT (Devlin et al., 2019) and XLM-Roberta (Conneau et al., 2020) are the state-of-the-art for cross-lingual transfer learning on natu-

| | EN | ES | DE | FR |
|---|---|---|---|---|
| XLM-Roberta | 84.7 | 79.4 | 77.4 | 79.1 |
| MMTE | 79.6 | 71.6 | 68.2 | 69.5 |
| MTL | 84.8 | 80.1 | 78.3 | 79.8 |

Table 7: Evaluation on XNLI task, XLM-Roberta results is our reproduction of the results. Massively Multilingual Translation Encoder (MMTE) is reported from (Siddhant et al., 2020)

ral language understanding (NLU) tasks, such as XNLI (Conneau et al., 2018) and XGLUE (Liang et al., 2020). Such models are trained on massive amount of monolingual data from all language as a masked language model. It has been shown that massive MNMT models are not able to match the performance of pre-trained language models such as XLM-Roberta on NLU downstream tasks (Siddhant et al., 2020). In Siddhant et al. (2020), the MNMT models are massive scale models trained only on the NMT task. They are not able to outperform XLM-Roberta, which is trained with MLM task without any parallel data. In this work, we evaluate the effectiveness of our proposed MTL approach for cross-lingual transfer leaning on NLU application. Intuitively, MTL can bridge this gap since it utilizes NMT, MLM and DAE objectives.

In the experiment, we train a system on 6 languages using both bitext and monolingual data. For the bitext training data, we use 30M parallel sentences per language pair from in-house data crawled from the web. For the monolingual data, we use 40M sentences per language from CC-Net (Wenzek et al., 2019). Though this is a relatively large-scale setup, it only leverages a fraction of the data used to train XLM-Roberta for those languages. We train the model with 12 layers encoders and 6 layers decoder. The hidden dimension is 768 and the number of heads is 8. We tokenize all data with the SentencePiece model (Kudo and Richardson, 2018) with the vocabulary size of 64K. We train a many-to-many MNMT system with three tasks described in Section 3.1: NMT, MLM, and DAE. Once the model is trained, we use the encoder only and discard the decoder. We add a feedforward layer for the downstream tasks.

As shown in Table 7, MTL outperform both XLM-Roberta and MMTE (Siddhant et al., 2020) which are trained on massive amount of data in comparison to our system. XLM-Roberta is trained only on MLM task and MMTE is trained only on NMT task. Our MTL system is trained on three tasks. The results clearly highlight the effectiveness of multi-task learning, and demonstrate that it can outperform single-task systems trained on massive amount of data. We observe the same pattern in Table 8 with XGLUE NER task, which outperforms SOTA XLM-Roberta model.

| | EN | ES | DE |
|---|---|---|---|
| XLM-Roberta | 91.1 | 76.5 | 70.3 |
| MTL | 91.2 | 77.0 | 75.0 |

Table 8: Evaluation on XGLUE NER task, XLM-Roberta results is our reproduction of the results.

## 6 Conclusion

In this work, we propose a multi-task learning framework that jointly trains the model with the translation task on bitext data, the masked language modeling task on the source-side monolingual data and the denoising auto-encoding task on the target-side monolingual data. We explore data and noising scheduling approaches and demonstrate their efficacy for the proposed approach. We show that the proposed MTL approach can effectively improve the performance of MNMT on both high-resource and low-resource languages with large margin, and can also significantly improve the translation quality for zero-shot language pairs without bitext training data. We showed that the proposed approach is more effective than pre-training followed by fine-tuning for NMT. Furthermore, we showed the effectiveness of multitask learning for cross-lingual downstream tasks outperforming SOTA larger models trained on single task.

For future work, we are interested in investigating the proposed approach in a scaled setting with more languages and a larger amount of monolingual data. Scheduling the different tasks and different types of data would be an interesting problem. Furthermore, we would also like to explore the most sample efficient strategy to add a new language to a trained MNMT system.

## Acknowledgment

We would like to thank Alex Muzio for helping with zero-shot scoring and useful discussions. We also would like to thank Dongdong Zhang for helping with mBART comparison and Felipe Cruz Salinas for helping with XNLI and XGLUE scoring.

## References


Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 1538–1548.

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.

Rich Caruana. 1997. Multitask learning. *Machine learning*, pages 41–75.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Adrià De Gispert and Jose B Marino. 2006. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *Proc. of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 65–68. Citeseer.

Li Deng, Geoffrey Hinton, and Brian Kingsbury. 2013. New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

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.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, pages 71–99.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. 2017. Learning to parse and translate improves neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 72–78.

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.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1243–1252. JMLR. org.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 344–354.

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

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

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*, pages 339–351.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

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

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

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*, pages 5039–5049.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 296–301.

Jan Niehues and Eunah Cho. 2017. Exploiting linguistic resources for neural machine translation using multi-task learning. In *Proceedings of the Second Conference on Machine Translation*, pages 80–89.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Conference on Machine Translation*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised nmt using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8854–8861. AAAI Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2062–2068.

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*.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103.

Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2019. Multi-agent dual learning. In *International Conference on Learning Representations*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzman, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Lijun Wu, Yiren Wang, Yingce Xia, QIN Tao, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for 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 4198–4207.

Poorya Zaremoodi and Gholamreza Haffari. 2018. Neural machine translation for bilingually scarce scenarios: a deep multi-task learning approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.

# Appendices

## A Bitext Training Data

We concatenate all resources except WikiTitles provided by WMT of the latest available year and filter out duplicated pairs and pairs with the same source and target sentence. For Fr and Cs, we randomly sample 10M sentence pairs from the full corpus. The detailed statistics of bitext data can be found in Table 9.

We randomly sample $1,000$ sentence pairs from each individual validation set and concatenate them to construct a multilingual validation set. We tokenize all data with the SentencePiece model (Kudo and Richardson, 2018), forming a vocabulary shared by all the source and target languages with 32k tokens for bilingual models (16k for Hi and Gu) and 64k tokens for multilingual models.

| Code | Language | #Bitext | Validation |
|---|---|---|---|
| Fr | French | 10M | Newstest13 |
| Cs | Czech | 10M | Newstest16 |
| De | German | 4.6M | Newstest16 |
| Fi | Finnish | 4.8M | Newstest16 |
| Lv | Latvian | 1.4M | Newsdev17 |
| Et | Estonian | 0.7M | Newsdev18 |
| Ro | Romanian | 0.5M | Newsdev16 |
| Hi | Hindi | 0.26M | Newsdev14 |
| Tr | Turkish | 0.18M | Newstest16 |
| Gu | Gujarati | 0.08M | Newsdev19 |

Table 9: Statistics of the parallel resources from WMT. A list of 10 languages ranked with the size of bitext corpus translating to/from English.