Gradient Consistency-based Parameter Allocation for Multilingual Neural Machine Translation

Wenshuai Huo[1,2], **Xiaocheng Feng**[*1,2] , **Yichong Huang**[1], **Chengpeng Fu**[1,2], **Hui Wang**[2], **Bing Qin**[*1,2]
[1] Harbin Institution of Technology [2] Peng Cheng Laboratory
{wshuo, xcfeng, ychuang, cpfu, qinb}@ir.hit.edu.cn
wangh06@pcl.ac.cn

Abstract

Multilingual neural machine translation handles the translation of multiple languages with one unified model. However, this joint-training paradigm incurs the notorious issue of *parameter interference*, where the model compromises with the language diversity to find a common solution. Recent research has explored avoiding this problem by selecting certain parameters for each language direction from the original model to form language-specific sub-networks. However, determining how many parameters to choose and which parameters to select is still a serious challenge. In this work, we propose an approach called CaPA (**C**onsistency-b**a**sed **P**arameter **A**llocation), which dynamically allocates parameters of appropriate scale to each language direction based on the consistency between the gradient of the individual language and the average gradient. Specifically, CaPA allocates more parameters to languages with higher gradient consistency as these languages tend to have a more positive impact on other languages. Furthermore, considering the varying levels of interference across different parts of the model, we propose an adaptive parameter allocation based on module-level gradient consistency. Experimental results show the correlation between gradient consistency and parameter interference, as well as the effectiveness of our proposed method.

Keywords: Multilinguality, Machine Translation

1. Introduction

Neural machine translation (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Vaswani et al., 2017; Wang et al., 2019) has made great progress in recent years. Multilingual neural machine translation (MNMT) (Johnson et al., 2017; Tan et al., 2019; Fan et al., 2021; Huang et al., 2022b, 2023) is a variant of NMT that is designed to handle translation between multiple languages with one single model, which can be more efficient and cost-effective.

In the standard MNMT model, different languages share all of the model parameters. The paradigm of completely parameters sharing enables the transfer of cross-lingual knowledge, significantly enhancing the translation performance of low-resource languages. However, it also gives rise to the issue of parameter interference, i.e., different languages express a disagreement on some parameters, which leads to the model a compromise across all languages. This compromise may not capture the specific nuances of individual languages, leading to reduced translation quality. In order to learn language specific knowledge, researchers have explored various strategies. These include the incorporation of additional language-specific components, such as the language-specific attention (Blackwood et al., 2018), adapter (Bapna and Firat, 2019; Baziotis et al., 2022), layer (Zhang et al., 2020; Pires et al., 2023) or the duplication of parameters that give rise to conflicts among different language pairs (Wang and Zhang, 2022). However, it is worth noting that these approaches inevitably lead to a notable increase in the overall model parameters.

Another line of research in language-specific modeling aims to precisely specify certain parameters in the original model through the utilization of model pruning methodologies (Lin et al., 2021; Xie et al., 2021). The specified parameters for each language direction form language-specific sub-networks, and each sub-network can be divided into two parts: shared parameters, which retain general knowledge, and language-specific parameters, which capture language-specific knowledge. However, determining how many parameters for each language direction and selecting which parameters to include still remains a challenging task. Lin et al. (2021) manually set a uniform percentage for all language directions, which is both cumbersome and limited in the performance. On the other hand, traditional methods for learning sub-networks solely assign parameters based on their importance within individual languages, disregarding the relevance between languages.

In this work, we propose a novel pruning method for multilingual machine translation that adaptively allocates appropriately sized sub-networks for each language direction. We define the similarity between the gradient of each language direction and

*Corresponding author

LREC-COLING 2024, pages 7901–7912
20-25 May, 2024.

the one of the overall training objective as gradient consistency, and the amount of parameters allocated to each language direction is determined based on consistency. Therefore, language directions with higher inter-language consistency, which tend to positively contribute to the overall objective, can be allocated more parameters. On the contrary, if the language exhibits a low level of gradient consistency, it indicates that this language will cause interference to the overall performance. In such cases, the impact of this interference should be alleviated by reducing the allocated parameter quantity. Furthermore, the level of interference in different parts of the model often varies. So we propose to pertinently allocate parameters based the module level gradient consistency.

Our contributions can be summarized as follows:

- We propose an innovative pruning method for multilingual machine translation that dynamically allocates parameters of appropriate scales to each language direction based on the gradient consistency among different language directions.
- We propose an adaptive parameter allocation based on module-level consistency to effectively alleviate interference and retain more general knowledge.
- Our method is parameter-efficient and does not significantly increase the number of parameters.
- Experimental results demonstrate that our method facilitates the sharing of general knowledge while suppresses the cross-lingual negative transfer[1].

2. Background

Multilingual Neural Machine Translation The standard paradigm of MNMT is to use a single model with completely shared parameters to translate L language pairs $\{S^1 - T^1, S^2 - T^2, ..., S^L - T^L\}$, from any source language S^l to its traget T^l. To indicate the translation direction, special language token are added at the beginning of source and target sentences. In this work, we follow Johnson et al. (2017) to employ the Transformer as the backbone network.

Learning Language Specific Sub-networks Learning language-specific sub-networks can mitigate the interference between languages in multilingual machine translation model. The shared parameters tend to retain general knowledge, while the language-specific parameters focus on learning language-specific details. In this work, we follow Lin et al. (2021) and selectively allocate parameters using the method of model pruning. A sub-network for language pair l is indicated by a binary mask vector $\mathbf{M}^l \in \{0,1\}^\theta$ in base model θ. Each element indicates whether the weight w is retained for language pair l. During training and inference, only the parameters belonging to sub-network l are involved in the computations, rather than the entire model.

[1] https://github.com/huowsh/CaPA

Algorithm 1 Gradient Consistency-based Parameter Allocation

Input: *language pairs number* L; *training data* $\{D_l^{train}\}_{l=1}^L$; *development data* $\{D_l^{dev}\}_{l=1}^L$;
Initialize: MNMT model θ; fine-tune θ for $l \in [1, L]$;

while θ not converge **do**
  train θ *for multiple steps with* D^{train}
  for $l \in [1, L]$ **do**
    sample a batch of data B_l^{dev} *from* D_l^{dev}
    calculate gradient $grad^l$ *on* B_l^{dev}
  end for
  for $l \in [1, L]$ **do**
    $grad^{avg} = \sum_{i=1, i\neq l}^{L} grad^i$
    $consistency = \cos(grad^l, grad^{avg})$
    reallocate a sub-network for l
  end for
end while

3. Approach

In this section, we describe the proposed parameter allocation approach. We first introduce the overall process of CaPA. And we compare the difference between our method and previous work on learning language specific sub-networks. Then, we provide a detailed description of CaPA, which dynamically allocates parameters for each language direction based on the gradient consistency observed during the training process.

Furthermore, we found that the gradient consistency between translation directions varies across different parts of the network. Therefore, we propose a parameter allocation scheme based on module-level gradient consistency.

3.1. Overall

Algorithm 1 lists the overall process of our method. Initially, we train an entire parameter-shared multilingual translation model θ. Subsequently, similar to the approach in LaSS (Lin et al., 2021), we fine-tune the model θ on each language direction to identify the language-specific important weights based on their magnitudes. Different from their method, we proceed to train θ while adaptively allocating varying sizes of sub-networks to each language direction, taking into account the degree of gradient consistency.

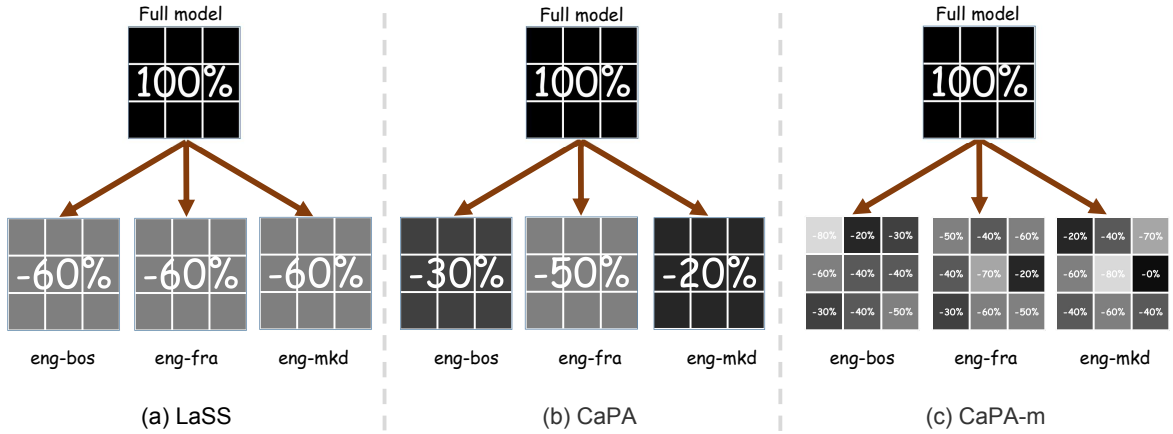

Figure 1: The illustration of LaSS (a) and the proposed methods CaPA (b) and CaPA-m (c) in learning language sub-network. (a) prunes the model weights by a uniform percentage R for all language pairs to obtain language-specific sub-networks, where R is a manually set hyper-parameter. (b) allocates different percentages of weights to each language based on the level of gradient consistency. (c) further evaluates the gradient consistency at the module level and only prunes the weights in the modules that exhibit conflict.

3.2. Gradient Consistency-based Parameter Allocation

As shown in (a) and (b) of Figure 1, unlike Lin et al. (2021) which prunes the weights at a fixed rate for all language directions, CaPA observes the gradient consistency during the training process and allocate different size of sub-networks to each language direction based on the degree of interference. Our main idea is that when the update objective for a language-pair significantly deviates from the average one, it means that this language is negatively interfering with the overall goal and should allocate fewer parameters. On the other hand, if the update targets are more similar, it means that the language is promoting the overall goal and should have more parameters. We define the similarity between the gradient of language pair l and the average gradient as *consistency* $\mathcal{C}^l$:

$$\mathcal{C}^l = \cos(grad^l, grad^{avg}), \tag{1}$$

the greater the $\mathcal{C}^l$, the more l can promote the overall training goal. On the contrary, it indicates that l will cause interference and more parameters should be pruned for l. So the pruning rate R of l can be defined as:

$$R^l = \sigma\big(1 - \frac{2}{1 + \exp\big(-\beta(\mathcal{C}^l - \lambda)\big)}\big), \tag{2}$$

where R^l and $\mathcal{C}^l$ are negatively correlated; the hyperparameters λ and β respectively determine the starting point of pruning and the sensitivity of pruning rate to the influence of $\mathcal{C}^l$; σ represents the taking only of values greater than 0. When $(\mathcal{C}^l - \lambda)$ is greater than 0, it indicates a strong correlation between l and others . Therefore, we set $R^l = 0$. And when $(\mathcal{C}^l - \lambda)$ is less than 0, the range of values for R^l is between $(0, 1)$, and the smaller the value of $\mathcal{C}^l$ the fewer parameters are allocated to l with a bigger R^l.

Considering that the gradient similarity situation will change during training, we will dynamically update the pruning rate. The method for calculating the pruning rate R^l_t of language l at time t is as follows:

$$R^l_t = \sigma\big((1 - \frac{2}{1 + \exp\big(-\beta(\mathcal{C}^l - \lambda)\big)}) (1 - R^l_{t-1}) + R^l_{t-1}\big), \tag{3}$$

R^l_t is calculated based on R^l_{t-1} from the previous iteration. $(1 - R^l_{t-1})$ represents the ratio of the parameters that are still related to l after the previous pruning to the total number of parameters in the model. The new sub-network is formed by merging the required changes with the previous sub-network.

3.3. Module-level Gradient Consistency-based Parameter Allocation

Through Section 3.2, we calculate the pruning rate based on the *consistency* of each language pair on the entire model. And similar to LaSS(Lin et al., 2021), we apply the R^l to each module in the network, as illustrated in (a) and (b) of Figure 1. However, the degree of interference in different parts of the model is often not the same. Therefore, we can calculate the *consistency* of each language on each component of the model. Formally, the *consistency* of the language pair l on module m in

TED-8-DIVERSE		TED-8-RELATED		WMT	
language	#num	language	#num	language	#num
bos (Bosnian)	5,664	bel (Belarusian)	4,509	tr (Turkish)	5,000
mar (Marathi)	9,840	aze (Azerbaijani)	5,946	ro (Romanian))	10,000
hin (Hindi)	18,798	glg (Glacian)	10,017	et (Estonian)	80,000
mkd (Macedonian)	25,335	slk (Slovak)	61,470	zh (Chinese))	400,000
ell (Greek)	134,327	cse (Czech)	103,093	de (German)	1,500,000
bul (Bulgarian)	174,444	tur (Turkish)	182,470	fr (French)	3,000,000
fra (French)	192,304	por (Portuguese)	184,755		
kor (Korean)	205,640	rus (Russian)	208,458		

Table 1: Data statistics for the TED-8-Diverse dataset, the TED-8-Related dataset and the WMT dataset. '#num' refers to the number of sentence pairs in the training set.

the MNMT model is defined by:

$$\mathcal{C}^{l_m} = \cos(grad^{l_m}, grad^{avg_m}), \tag{4}$$

$grad^{l_m}$ represents the backpropagated gradient for language direction l on module m, while $grad^{avg_m}$ denotes the average gradient of all language directions on module m. Then we will obtain the adaptive pruning rate of each module based on the local *consistency*:

$$R_t^{l_m} = \sigma\big((1 - \frac{2}{1+\exp\big(-\beta(\mathcal{C}^{l_m} - \lambda)\big)}) (1 - R_{t-1}^{l_m}) + R_{t-1}^{l_m}\big). \tag{5}$$

So, we can distinguish whether each module has interference and prune according to the degree of interference as (c) in Figure 1. While more pertinently relieving interference, retain more general knowledge.

4. Experiments

4.1. Settings

Datasets We conducted experiments on publicly available multilingual datasets, including the widely-used benchmark TED-8-Diverse and TED-8-Related (Wang et al., 2020a), and a relative large-scale WMT dataset(Bojar et al., 2014, 2016, 2017, 2018) .

The TED-8-diverse dataset contains 4 low-resource languages (bos, mar, hin, mkd) and 4 high-resource languages (ell, bul, fra, kor). The TED-8-Related contains 4 low-resource languages (aze, bel, glg, slk) and 4 related high-resource language (tur, rus, por, ces). We follow Wang et al. (2020a) to apply sentencepiec (Kudo and Richardson, 2018) to preprocess sentences with vocabulary sizes of 8k for both TED-8-related and TED-8-diverse datasets.

For the WMT dataset, it includes 3 low-resource languages (et, ro, tr) and 3 high-resource languages (fr, de, zh) to English from WMT14, WMT16, WMT17, and WMT18.

In each dataset, we conduct experiments for two multilingual translation tasks: 1) Many-to-One (M2O), where multiple languages are translated into English; 2) One-to-Many (O2M), where English is translated into various other languages. The details of the datasets are listed in Table 1

Model Settings In this work, we follow Johnson et al. (2017) to employ the transformer as the backbone network. In order to maintain consistency with previous work, we perform our experiments with variants of transformer architecture for different datasets. For TED-8-Diverse dataset and the TED-8-Related dataset, we adopt the transformer base setting which includes 6 encoder and decoder layers, 512/1024 hidden dimensions and 4 attention heads. And for the larger dataset WMT, the architecture we adopt includes 6 encoder and decoder layers, 512/2048 hidden dimensions and 8 attention heads. All the translation models are implemented with the fairseq-py[2] (Ott et al., 2019). The following are the hyperparameters details:

- We use Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$, and the learning rate is set to 0.0005.
- The values of β and λ are set to 50 and 0.03, respectively.
- Before commencing the parameter reallocation, we conduct a warm-up training of 1000 steps.
- The interval for calculating gradient consistency and reassigning subnetworks is one epoch
- The batch size is configured as 64K.
- We apply dropout with a rate of 0.1.
- We employ temperature-based sampling with a temperature parameter (τ) of 1.
- We employ beam search with a beam width of 5.

[2] https://github.com/facebookresearch/fairseq.

Method	TED-8-DIVERSE O2M	TED-8-DIVERSE M2O	TED-8-RELATED O2M	TED-8-RELATED M2O	WMT O2M	WMT M2O
Baselines						
MULTILINGUAL	20.67	26.81	19.42	25.29	19.41	20.17
LaSS (Lin et al., 2021)	21.2	27.14	19.61	25.9	19.86	19.7
MultiDDS-S (Wang et al., 2020a)*	18.24	27.00	17.32	25.52	–	–
LaGMD (Pham et al., 2022)*	20.1	27.7	19.3	26.3	–	–
Our Proposed Approaches						
CaPA	**21.81**	**27.92**	**20.31**	26.29	**20.36**	20.24
CaPA-m	21.64	27.88	20.07	**26.37**	20.34	**20.31**

Table 2: BLEU score on TED-8-Diverse, TED-8-Related and WMT datasets. Our method surpasses other multilingual baselines. CaPA and CaPA-m represent our parameter allocation method based on the overall gradient consistency and the module-level gradient consistency respectively. '*' represents results taken from original papers. **Bold** indicates the best performance.

TED-8-Diverse M2O	bos	mar	hin	mkd	ell	bul	fra	kor	Avg
LaSS	+0.9	-0.2	+0.09	+0.93	+0.28	+0.11	+0.32	+0.26	+0.33
CaPA	**+1.76**	**+0.39**	**+0.99**	**+1.28**	**+1.3**	**+1.18**	**+0.98**	**+1.09**	**+1.11**
TED-8-Diverse O2M	**bos**	**mar**	**hin**	**mkd**	**ell**	**bul**	**fra**	**kor**	**Avg**
LaSS	+0.48	+0.12	-0.03	+1.26	+0.83	+0.51	+0.79	+0.19	+0.53
CaPA	**+1.36**	**+0.63**	**+0.9**	**+1.65**	**+1.22**	**+1.4**	**+1.36**	**+0.55**	**+1.14**
TED-8-Related M2O	**aze**	**bel**	**glg**	**slk**	**tur**	**rus**	**por**	**ces**	**Avg**
LaSS	**+0.24**	+0.45	+0.95	+0.89	+0.76	+0.47	+0.52	+0.7	+0.61
CaPA	+0.22	**+0.88**	**+1.26**	**+1.29**	**+1.3**	**+0.71**	**+1.14**	**+1.16**	**+1.0**
TED-8-Related O2M	**aze**	**bel**	**glg**	**slk**	**tur**	**rus**	**por**	**ces**	**Avg**
LaSS	-0.08	-0.3	+0.08	+0.08	+0.4	+0.43	+0.53	+0.35	+0.19
CaPA	**+0.26**	**+0.3**	**+0.5**	**+0.88**	**+1.64**	**+1.23**	**+1.26**	**+1.07**	**+0.89**

Table 3: BLEU score improvements of LaSS and our CaPA over the MULTILINGUAL baseline. CaPA outperforms the MULTILINGUAL baseline in every language direction. **Bold** indicates the best performance. Languages are ordered increasingly by data size from left to right.

- We detokenize the final translation results and evaluate the quality with 4-gram BLEU (Papineni et al., 2002) score by SacreBLEU [3](Post, 2018).

Baselines We compare our methods with: 1) MULTILINGUAL, the standard paradigm of MNMT (Johnson et al., 2017); 2) LaSS(Lin et al., 2021), the most related study to our approach, learns language-specific sub-networks based solely on the importance of parameters. To re-implement LaSS, we tried pruning rates from 0.1 to 0.9 to partition the language-specific sub-network for all language directions; 3) MultiDDS-S (Wang et al., 2020a) utilizes gradient similarity to dynamically adjust the sampling rate for each language direction, achieving a balanced training for MNMT; 4) LaMGD (Pham et al., 2022), which also aims to map sub-networks to language directions by masking the output of each layer.

[3]https://github.com/mjpost/sacrebleu

4.2. Main Results

The experimental results in Table 2 demonstrate the effectiveness of our methods. It shows that our approaches outperform all baselines on the TED-8-Diverse and TED-8-related datasets. Compared to LaSS(Lin et al., 2021), our approach exhibits superior performance on both the M2O and O2M tasks, indicating its enhanced ability to facilitate positive transfer among different language directions while suppressing negative transfer.

For the WMT dataset, on the M2O translation task, LaSS did not surpass the baseline, and the improvements achieved by our method were also limited. This could be attributed to the phenomenon of parameter interference is not severe in the M2O task(Zhu et al., 2021; Shaham et al., 2022) in multilingual machine translation. Our method, which

	→ de	→ en	→ it	→ nl	→ ro
de →	–	34.62 / **35.39**	21.82 / **22.64**	23.39 / **23.71**	18.52 / **18.64**
en →	28.68 / **29.64**	–	29.96 / **30.32**	29.84 / **30.49**	26.05 / **26.48**
it →	20.72 / **21.03**	33.74 / **34.03**	–	21.38 / **22.22**	19.44 / **19.80**
nl →	22.36 / **22.76**	33.84 / **34.59**	20.99 / **21.35**	–	18.45 / **18.85**
ro →	22.09 / **22.43**	36.06 / **36.72**	23.84 / **24.06**	22.02 / **22.10**	–

Table 4: The M2M translation results on the IWSLT dataset. The BLEU scores are reported in the format of MULTILINGUAL/Ours. **Bold** indicates the better result. CaPA outperforms the MULTILINGUAL baseline in every language direction.

is oriented to address the cross-lingual negative interference problem, is more valuable for O2M. A more in-depth analysis of this is provided in Section 5.1.

The module-level CaPA achieves better results than the baseline; however, there is no significant improvement relative to CaPA. This observation suggests that, in contrast to the overall gradient consistency differences between language pairs, the module-level gradient consistency differences may have a relatively minor impact. Further analysis is provided in Section 5.4.

4.3. Results On Each Language Direction

We calculated the differences between CaPA and LaSS results compared to the MULTILINGUAL baseline in each language direction, as shown in Table 3. CaPA outperforms the MULTILINGUAL baseline in every language direction.

Both CaPA and LaSS tend to achieve more significant improvements in high-resource languages. This is because in multilingual machine translation, there's often a trade-off where performance in high-resource language directions is sacrificed to boost low-resource directions, resulting in more severe cross-lingual interference challenges in high-resource directions(Aharoni et al., 2019; Yang et al., 2022; Shaham et al., 2022). LaSS and our method, both of which can mitigate the issue of cross-lingual negative interference, lead to more significant improvements in high-resource directions. This is a reasonable outcome.

While LaSS performs well in high-resource language directions, it suffers from diminished effectiveness in low-resource language directions. CaPA addresses this issue.

4.4. M2M Results

We conducted experiments on the IWSLT17(Cettolo et al., 2017) dataset for the many-to-many (M2M) scenario, which encompasses data for all 20 language directions of 5 languages (`en`, `it`, `de`, `nl`, `ro`). The results are shown in Table 4. Our method outperforms the baseline in all 20 translation directions.

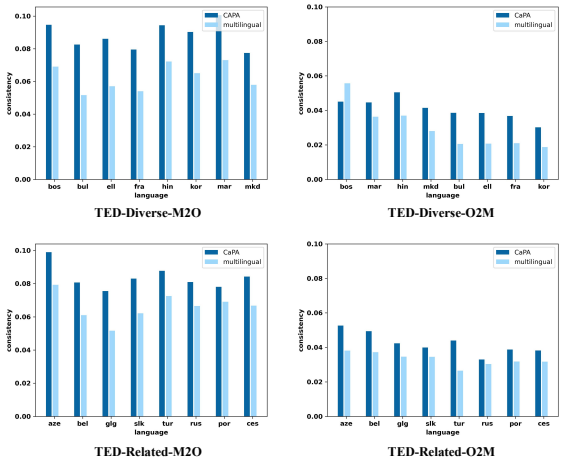

Figure 2: The gradient consistency on each language direction during the training process for CaPA and the baseline. CaPA exhibits superior performance compared to the baseline in almost all language directions. The results are derived from the TED-8-Diverse dataset and the TED-8-Related dataset.

5. Analysis

5.1. Consistency Among Languages

We extracted the average proficiency level of *consistency* from CaPA and MULTILINGUAL throughout the training process, until reaching optimal performance, as depicted in Figure 2. CaPA demonstrates superior performance compared to the baseline in almost all language directions. This indicates that CaPA is effective in suppressing cross-lingual negative interference and promoting positive cross-lingual knowledge transfer.

In multilingual machine translation, the negative interference problem is more severe in the One-to-Many (O2M) task (Zhu et al., 2021; Shaham et al., 2022). This is due to O2M requiring the decoding of text into multiple languages, which necessitates considering differences in vocabulary, syntax, style, and more in the target language. Observing Figure 2, it becomes apparent that whether we consider the baseline or our proposed approach, M2O consistently exhibits higher gradient consis-

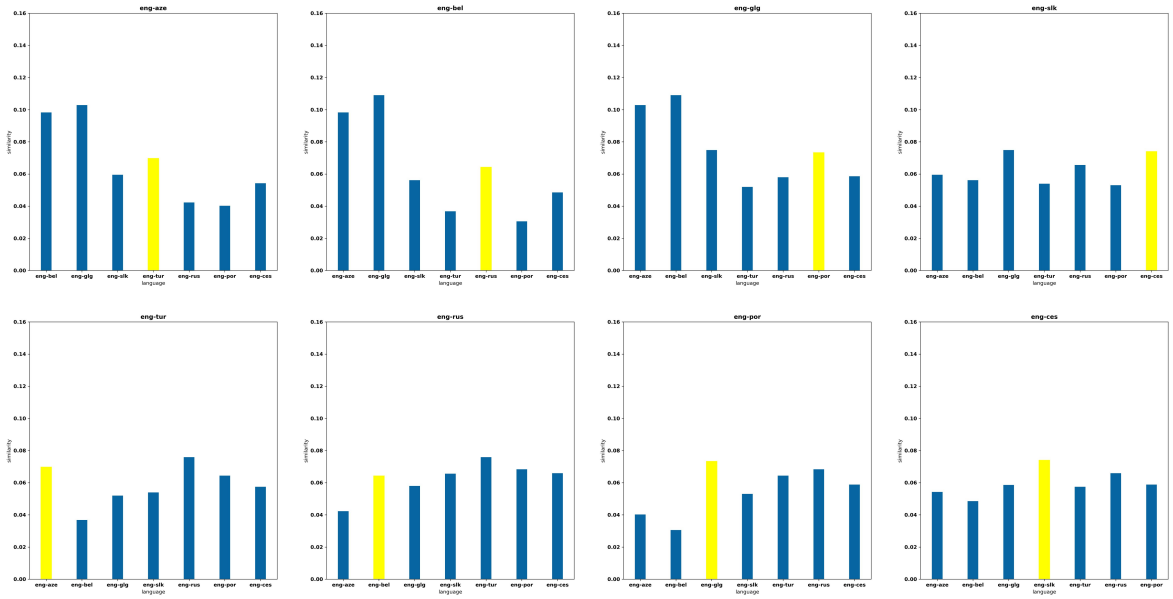

Figure 3: Gradient similarity between each language direction in TED-8-Related for the O2M translation task. When dataset scales align, there is a correlation between gradient similarity and language similarity. The yellow columns highlight languages that belong to the same language family. Languages are ordered increasingly by data size from left to right.

tency across various language directions. This indicates a correlation between gradient consistency and cross-lingual interference. This finding also suggest that leveraging the degree of gradient consistency may play a pivotal role in mitigating the challenges posed by cross-lingual interference.

5.2. The Relationship Between The Data Size And Gradient consistency

In view of Figure 2, we observe that gradient consistency is generally higher in low-resource directions, especially in the O2M direction.

In multilingual machine translation, high-resource language directions are more susceptible to severe cross-lingual interference challenges(Aharoni et al., 2019; Yang et al., 2022; Shaham et al., 2022). We speculate that this may be due to the lower translation performance in low-resource directions, which tends to learn more general knowledge. In contrast, high-resource language directions need to learn more complex language-specific knowledge. The lower gradient consistency in high-resource language directions also offers evidence of a correlation between gradient consistency and cross-lingual interference.

5.3. The Relationship Between Language Similarity And Gradient Similarity

In multilingual machine translation, language similarity is an intuitively important factor to consider(Lin et al., 2019; Wang et al., 2020c; Chronopoulou et al., 2023). Therefore, we are investigating the correlation between language relatedness and gradient similarity.

The TED-Related-8 dataset includes four sets of related languages: aze and tur (Turkic), bel and rus (Slavic), glg and por (Romance), ces and slk (Czech-Slovak). We calculate the pairwise gradient similarity between each language direction in the dataset during the training process. We present the result on TED-Related-8 O2M in Figure 3. In each bar chart, every column represents the gradient similarity between a specific language and all the other languages in the dataset. From left to right, the dataset scales incrementally. The yellow columns highlight languages that belong to the same language family.

We observed that the gradient similarity between similar languages is not consistently the highest but frequently exceeds that of adjacent columns. This observation suggests that as the dataset scales align, a correlation emerges between gradient similarity and language relatedness. Furthermore, the result also highlights the influence of dataset scale on gradient similarity.

5.4. Analysis Of The Results For CaPA And CaPA-m

Our method in Section 3.3 learning sparse subnetworks based on module-level interference. Theoretically, CaPA-m should outperform CaPA due to stronger flexibility. Unfortunately, this doesn't

TED-8-Diverse O2M	bos	mar	hin	mkd	ell	bul	fra	kor
CaPA	5.6%	3.7%	5.4%	3.2%	6.2%	4.6%	8.2%	5.1%
CaPA-m	0~18.1%	0~16.1%	0~15.1%	0~17.0%	0~18.9%	0~18.7%	0~20.6%	0~19.2%
TED-8-Diverse M2O	**bos**	**mar**	**hin**	**mkd**	**ell**	**bul**	**fra**	**kor**
CaPA	0.1%	0.1%	0.1%	1.5%	2.0%	1.9%	1.4%	1.0%
CaPA-m	0~9.2%	0~11.3%	0~7.1%	0~9.9%	0~7.4%	0~9.8%	0~11.3%	0~14.1%
TED-8-Related O2M	**aze**	**bel**	**glg**	**slk**	**tur**	**rus**	**por**	**ces**
CaPA	5.8%	7.7%	5.2%	3.2%	6.9%	9.5%	5.6%	5.0%
CaPA-m	0~18.0%	0~17.6%	0~23.6%	0~16.9%	0~16.2%	0~18.2%	0~18.5%	0~18.3%
TED-8-Related M2O	**aze**	**bel**	**glg**	**slk**	**tur**	**rus**	**por**	**ces**
CaPA	0.1%	0.2%	0.2%	0.6%	1.8%	1.4%	1.7%	2.1%
CaPA-m	0~7.8%	0~8.2%	0~9.5%	0~9.0%	0~3.6%	0~5.4%	0~3.6%	0~3.6%

Table 5: The average magnitude of parameter count changes caused by our method during the training process in the TED-8-Diverse and the TED-8-Related datasets.

always hold. We aim to evaluate whether CaPA-m brings about improvements in gradient consistency across every module. Due to space constraints, we are unable to provide details for each module. Therefore, we have computed the average consistency across all modules within each layer to determine the cross-layer distribution of consistency, as presented in Figure 4. We observed that CaPA-m exhibits an improvement in gradient consistency when compared to CaPA. However, this enhancement does not consistently align with the experimental outcomes in Section 4.2. This suggests that, apart from consistency, there are other factors at play influencing translation performance.

Drastic adjustments in model scale might lead to the anomaly. Given that our approach dynamically adjusts the scale of subnetworks throughout training, a sudden change (increasing or decreasing) in the number of parameters during training can introduce instability. This instability does not significantly impact performance when the magnitude of change is minor; however, it may damage performance when the magnitude of change is intense. We calculated the average magnitude of parameter count changes caused by our method during the training process. The results are presented in Table 5. For CaPA, this value remains the same across each module, while for CaPA-m, it varies across modules. We observed that for the O2M task, CaPA-m leads to significant parameter scale changes in certain modules. For the Related M2O dataset, even though CaPA-m also induces larger parameter scale changes compared to CaPA, the actual values are not substantial. This also explains why CaPA-m might perform better on the M2O task. Additionally, it's worth mentioning that the relatively higher magnitude of parameter count changes in O2M confirm the more severe negative interference in O2M compared to M2O, as mentioned in Section 5.1.

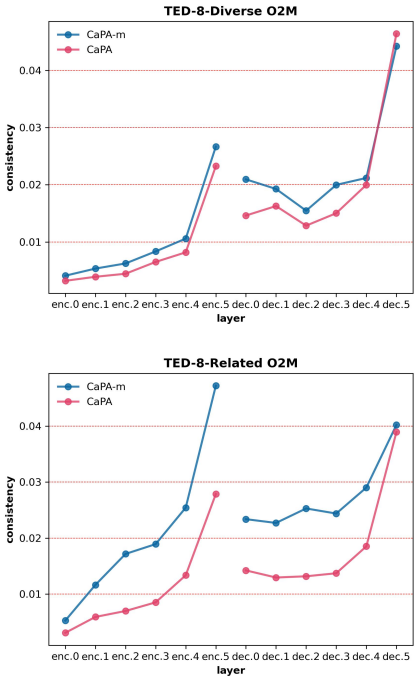

Figure 4: Average gradient consistency at each layer on the TED-8-Diverse and TED-8-Related datasets for the O2M task. CaPA-m exhibits an improvement in gradient consistency when compared to CaPA.

6. Related Work

Multilingual neural machine translation can facilitate cross-lingual knowledge transfer through parameter sharing, but it also lead to interference issue. To retain beneficial transfer and avoid negative interference, previous research has made significant efforts. Blackwood et al. (2018); Bapna

and Firat (2019); Zhang et al. (2020); Baziotis et al. (2022); Wang and Zhang (2022); Pires et al. (2023); Yuan et al. (2023) involve the language-specific components to learn language-specific knowledge. Arivazhagan et al. (2019b); Conneau et al. (2019); Wang et al. (2020a); Wu et al. (2021); Li and Gong (2021) address the issue of imbalanced data distribution in multilingual training. Arivazhagan et al. (2019a); Wei et al. (2020); Pan et al. (2021); Gu and Feng (2022); Gao et al. (2023) focus on promoting cross-lingual knowledge transfer by aligning semantic consistency representations.

Our work is also closely related to model pruning(See et al., 2016; Lan et al., 2019; Frankle and Carbin, 2018; Wang et al., 2020b), which typically remove unnecessary parameters from the model to compress its size and improve inference efficiency. Sun et al. (2020) first propose a sparsity-constrained parameter sharing mechanism that combines pruning methods for multi-task learning, Lin et al. (2021) and Xie et al. (2021) applied this strategy in the field of multilingual neural machine translation. It is noteworthy that the methods proposed by Sun et al. (2020),Lin et al. (2021) and Xie et al. (2021) are aimed at enhancing performance rather than compressing the model.

7. Conclusion

In this paper, we propose an adaptive transfer of cross-lingual knowledge in a multilingual machine translation model by dynamically allocating parameters of different scales to each language direction based on the gradient consistency. The experiment proved that our method can effectively improve translation results by relieving interference and promoting positive cross-lingual knowledge transfer. This paper also demonstrates a correlation between gradient consistency and cross-lingual interference. Future research can continue to focus on enhancing gradient consistency to mitigate the challenges of cross-lingual interference in multilingual machine translation.

8. Limitations

CaPA allocates fewer parameters for language pairs that interfere with the overall goal, but when the size of the sub-network is too small, it may cause a significant decline in performance. Our method is unable to automatically detect this situation, so we set a threshold to control the minimum size of the sub-network. Additionally, similar to Lin et al. (2021), our work also requires multiple steps of training to sort the parameters by their importance for each language direction. When dealing with a large number of language pairs, it can be quite tedious. We hope to improve above issues in future work.

9. Acknowledgements

Bing Qin and Xiaocheng Feng are the co-corresponding authors of this work. We thank the anonymous reviewers for their insightful comments. This work was supported by the National Key R&D Program of China via grant No. 2021ZD0112905, National Natural Science Foundation of China (NSFC) via grant 62276078, the Key R&D Program of Heilongjiang via grant 2022ZX01A32, the International Cooperation Project of PCL, PCL2022D01 and the Fundamental Research Funds for the Central Universities (Grant No.HIT.OCEF.2023018).

10. Bibliographical 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.

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

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

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Christos Baziotis, Mikel Artetxe, James Cross, and Shruti Bhosale. 2022. Multilingual machine translation with hyper-adapters. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1170–1185, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. 2023. On the pareto front of multilingual neural machine translation. *arXiv preprint arXiv:2304.03216*.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-family adapters for low-resource multilingual neural machine translation. In *Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 59–72.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. Scaling laws for multilingual neural machine translation. *arXiv preprint arXiv:2302.09650*.

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*.

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. Improving zero-shot multilingual neural machine translation by leveraging cross-lingual consistency regularization. *arXiv preprint arXiv:2305.07310*.

Shuhao Gu and Yang Feng. 2022. Improving zero-shot multilingual translation with universal representations and cross-mappings. *arXiv preprint arXiv:2210.15851*.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network.

Yichong Huang, Xiaocheng Feng, Xinwei Geng, Baohang Li, and Bing Qin. 2023. Towards higher pareto frontier in multilingual machine translation. *arXiv preprint arXiv:2305.15718*.

Yichong Huang, Xiaocheng Feng, Xinwei Geng, and Bing Qin. 2022a. Omniknight: Multilingual neural machine translation with language-specific self-distillation. *arXiv preprint arXiv:2205.01620*.

Yichong Huang, Xiaocheng Feng, Xinwei Geng, and Bing Qin. 2022b. Unifying the convergences in multilingual neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6822–6835.

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.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Xian Li and Hongyu Gong. 2021. Robust optimization for multilingual translation with imbalanced data. *Advances in Neural Information Processing Systems*, 34:25086–25099.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani,

Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 293–305.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Minh-Quang Pham, François Yvon, and Josep Crego. 2022. Latent group dropout for multilingual and multidomain machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2469–2481, Seattle, United States. Association for Computational Linguistics.

Telmo Pessoa Pires, Robin M. Schmidt, Yi-Hsiu Liao, and Stephan Peitz. 2023. Learning language-specific layers for multilingual machine translation.

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, Brussels, Belgium. Association for Computational Linguistics.

Abigail See, Minh-Thang Luong, and Christopher D Manning. 2016. Compression of neural machine translation models via pruning. *arXiv preprint arXiv:1606.09274*.

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2022. Causes and cures for interference in multilingual translation. *arXiv preprint arXiv:2212.07530*.

Tianxiang Sun, Yunfan Shao, Xiaonan Li, Pengfei Liu, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Learning sparse sharing architectures for multiple tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8936–8943.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Mingxuan Wang, Jun Xie, Zhixing Tan, Jinsong Su, Deyi Xiong, and Lei Li. 2019. Towards linear time neural machine translation with capsule networks. 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 803–812, Hong Kong, China. Association for Computational Linguistics.

Qian Wang and Jiajun Zhang. 2022. Parameter differentiation based multilingual neural machine translation.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020a. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Yong Wang, Longyue Wang, Victor OK Li, and Zhaopeng Tu. 2020b. On the sparsity of neural machine translation models. *arXiv preprint arXiv:2010.02646*.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020c. On negative interference in multilingual models: Findings and a meta-learning treatment. *arXiv preprint arXiv:2010.03017*.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2020. On learning universal representations across languages. *arXiv preprint arXiv:2007.15960*.

Minghao Wu, Yitong Li, Meng Zhang, Liangyou Li, Gholamreza Haffari, and Qun Liu. 2021. Uncertainty-aware balancing for multilingual and

multi-domain neural machine translation training. *arXiv preprint arXiv:2109.02284*.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. Importance-based neuron allocation for multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5725–5737.

Jian Yang, Yuwei Yin, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2022. Hlt-mt: High-resource language-specific training for multilingual neural machine translation. *arXiv preprint arXiv:2207.04906*.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2023. Lego-mt: Learning detachable models for massively multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11518–11533.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2020. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. 2021. Counter-interference adapter for multilingual machine translation. *arXiv preprint arXiv:2104.08154*.

11. Language Resource References

Bojar, Ondřej and Buck, Christian and Federmann, Christian and Haddow, Barry and Koehn, Philipp and Leveling, Johannes and Monz, Christof and Pecina, Pavel and Post, Matt and Saint-Amand, Herve and others. 2014. *Findings of the 2014 workshop on statistical machine translation*. PID https://www.statmt.org/wmt14/.

Bojar, Ondřej and Chatterjee, Rajen and Federmann, Christian and Graham, Yvette and Haddow, Barry and Huang, Shujian and Huck, Matthias and Koehn, Philipp and Liu, Qun and Logacheva, Varvara and others. 2017. *Findings of the 2017 conference on machine translation (wmt17)*. Association for Computational Linguistics. PID https://www.statmt.org/wmt17/.

Bojar, Ondrej and Chatterjee, Rajen and Federmann, Christian and Graham, Yvette and Haddow, Barry and Huck, Matthias and Yepes, Antonio Jimeno and Koehn, Philipp and Logacheva, Varvara and Monz, Christof and others. 2016. *Findings of the 2016 conference on machine translation (wmt16)*. Association for Computational Linguistics. PID https://www.statmt.org/wmt16/.

Bojar, Ondřej and Federmann, Christian and Fishel, Mark and Graham, Yvette and Haddow, Barry and Huck, Matthias and Koehn, Philipp and Monz, Christof". 2018. *Findings of the 2018 conference on machine translation (wmt18)*. Association for Computational Linguistics. PID https://www.statmt.org/wmt18/.

Cettolo, Mauro and Federico, Marcello and Bentivogli, Luisa and Niehues, Jan and Stüker, Sebastian and Sudoh, Katsuitho and Yoshino, Koichiro and Federmann, Christian. 2017. *Overview of the iwslt 2017 evaluation campaign*. PID https://workshop2017.iwslt.org/.

Wang, Xinyi and Tsvetkov, Yulia and Neubig, Graham. 2020. *Balancing Training for Multilingual Neural Machine Translation*. Association for Computational Linguistics. PID https://www.ted.com/participate/translate.