Tencent's Multilingual Machine Translation System for WMT22 Large-Scale African Languages

Wenxiang Jiao Zhaopeng Tu Jiarui Li Wenxuan Wang Jen-tse Huang Shuming Shi
Tencent AI Lab
{joelwxjiao,zptu,ultrali,jwxwang,jentsehuang,shumingshi}@tencent.com

Abstract

This paper describes Tencent's multilingual machine translation systems for the WMT22 shared task on Large-Scale Machine Translation Evaluation for African Languages. We participated in the **constrained** translation track in which only the data and pretrained models provided by the organizer are allowed. The task is challenging due to three problems, including the absence of training data for some to-be-evaluated language pairs, the uneven optimization of language pairs caused by data imbalance, and the curse of multilinguality. To address these problems, we adopt data augmentation, distributionally robust optimization, and language family grouping, respectively, to develop our multilingual neural machine translation (MNMT) models. Our submissions won the **1st** place on the blind test sets in terms of the automatic evaluation metrics.[1]

1 Introduction

Multilingual neural machine translation (MNMT) aims to translate between multiple language pairs with a unified model (Johnson et al., 2017). It is appealing due to the model efficiency, easy deployment, and knowledge transfer between high resource languages and low resource languages. Hence, MNMT has attracted more and more attention from both academia and industry. To improve the performance of MNMT models, previous researchers have proposed various approaches on advanced model architectures (Sen et al., 2019; Zhang et al., 2021), training strategies (Wang et al., 2020a,b), and data utilization (Siddhant et al., 2020; Wang et al., 2022). In addition, industrial companies have released massive multilingual pretrained models (Tang et al., 2021) and large-scale multilingual translation models (Fan et al., 2021; Team et al., 2022) to facilitate translation among hundreds of languages. However, existing efforts on MNMT for African languages are not sufficient due to the lack of high quality and standardized evaluation benchmarks.

In this paper, we build a system integrating several advanced approaches for WMT22 Large-Scale Machine Translation Evaluation Task (Adelani et al., 2022), which involves a set of 24 African languages. We participated in the Constrained Translation track, where only the data provided by the organizer are allowed. This task is challenging due to three potential problems:

- The absence of training data for some to-be-evaluated language pairs;
- The uneven optimization of language pairs due to data imbalance;
- The curse of multilinguality in MNMT models caused by the hundreds of language pairs.

For the first problem, we adopt data augmentation techniques to construct synthetic data for the language pairs without parallel training data (§3.1). Specifically, we use back-translation (Sennrich et al., 2016) and self-training (Jiao et al., 2021), and attach a special tag to the synthetic side of the data. For the second issue, we utilize distributionally robust optimization (DRO) method (Oren et al., 2019; Zhou et al., 2021) to balance the optimization process for different translation directions (§3.2). For the third issue, we isolate the potential conflicts between language pairs by language family grouping and finetune a model for each language group (§3.3).

Experimental results show that our system can significantly improve the performance of vanilla MNMT models, from 15.50 to 17.95 BLEU points (§4.2). Extensive analysis suggests that data augmentation could be harmful to the translation performance if used for training the final models

[1]Codes, models, and detailed competition results are available at https://github.com/wxjiao/WMT2022-Large-Scale-African.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1049–1056
Abu Dhabi, December 7–8, 2022.

Table 1: Information of language groups and the corresponding language pairs. We include additional 36 language pairs (**bolded**) to help the long-tail languages.

Group	Language Pairs (73) **(117)**
ENGC	afr-eng,amh-eng,eng-fra,eng-fuv,eng-hau,eng-ibo,eng-kam,eng-kin,eng-lug,eng-luo,eng-nso,eng-nya,eng-orm,eng-sna,eng-som,eng-ssw,eng-swh,eng-tsn,eng-tso,eng-umb,eng-xho,eng-yor,eng-zul, (23) ,**eng-lin, eng-wol, (25)**
FRAC	fra-kin,fra-lin,fra-swh,fra-wol, (4) ,**amh-fra,fra-kam,fra-lug,fra-luo,fra-orm,fra-umb, (10)**
SSEA	**afr-nso**,**afr-sna**,afr-ssw,afr-tsn,afr-xho,afr-tso,afr-zul,nso-sna,nso-ssw,nso-tsn,nso-xho,nso-tso,nso-zul,sna-ssw,sna-tsn,sna-xho,sna-tso,sna-zul,ssw-tsn,ssw-xho,ssw-tso,ssw-zul,tsn-xho,tsn-tso,tsn-zul,tso-xho,tso-zul,xho-zul, (28)
HCEA	**amh-luo**,amh-orm,amh-som,amh-swh,luo-orm,luo-som,luo-swh,orm-som,orm-swh,som-swh, (10)
NGG	fuv-hau,fuv-ibo,fuv-yor,hau-ibo,hau-yor,ibo-yor, (6)
CA	**kin-lin**,kin-lug,kin-nya,kin-swh, **lin-lug**,**lin-nya**,**lin-swh**,lug-nya,lug-swh,nya-swh, (10)
OTHER	**fuv-kin,fuv-nya,fuv-som,fuv-zul,kam-nya,kam-sna,kam-som,kam-swh,kam-tso,kam-zul,kin-yor,lug-sna,lug-zul,luo-nya,luo-sna,luo-zul,nya-umb,nya-yor,sna-umb,sna-yor,som-wol,som-yor,swh-umb,swh-yor,tso-yor,umb-zul,xho-yor,yor-zul, (28)**

directly, due to the error-prone synthetic sentence pairs. Instead, we utilize the resulting MNMT models as pretrained models to further finetune on clean datasets for the final models. The DRO technique is very effective in improving the translation quality across all language pairs, particularly on the dominant languages (e.g., eng and fra), which also calls for an improved DRO to benefit more on other languages. As for language family grouping, it especially improves the translation quality on one-to-many translations, which demonstrates its effectiveness in alleviating the curse of multilinguality issue. Finally, our submission won the **1st** place in the official evaluation in terms of the automatic evaluation metrics.

2 Data

In this section, we present the details of our data preparation.

2.1 Language Pairs

We utilize all available datasets from the official website (including those from the Data Track participants)[2], which provide either monolingual or parallel sentences. According to the evaluation instruction, we group the language pairs into 7 groups, namely, English-Centric (ENGC), French-Centric (FRAC), South/South East Africa (SSEA), Horn of Africa and Central/East Africa (HCEA), Nigeria and Gulf of Guinea (NGG), Central Africa (CA), and Other related pairs (OTHER), to train the MNMT models. Details are listed in Table 1.

We consider three subsets of language pairs for training different models:

- **Base-146**: We train the TRANSF-DEEP (§4.1) models on the to-be-evaluated language pairs in the first 6 groups, as well as the English-French (i.e., eng-fra) pair. In total, there are 81 language pairs but only 73 of them are provided with bitext data, which cover 146 translation directions (i.e., including both forward and backward).

- **Large-234**: The main issues of **Base-146** are that, some to-be-evaluated language pairs (e.g., afr-nso) are missing in the training data and some languages are heavily long-tailed due to the imbalanced choice of language pairs. To alleviate these issues, we extend another 36 language pairs for the long-tail languages and construct synthetic data for all the language pairs in ENGC, SSEA, HCEA, NGG and CA, which enables the training on 234 translation directions. We use these language pairs to train the TRANSF-DWIDE (§4.1) models.

- **Eval-106**: The official evaluation includes 100 translation directions[3], which were notified at the later stage of the competition. We focus on these directions by finetuning the TRANSF-DWIDE models on these directions. To ensure the data amount of each language, we include all ENGC directions, making the final 106 directions.

[2] https://www.statmt.org/wmt22/large-scale-multilingual-translation-task.html

[3] https://docs.google.com/document/d/11NYyJpJ4nhNIllwmF5kjkqfkaaEzXNU-CCO5E64MRDU/edit

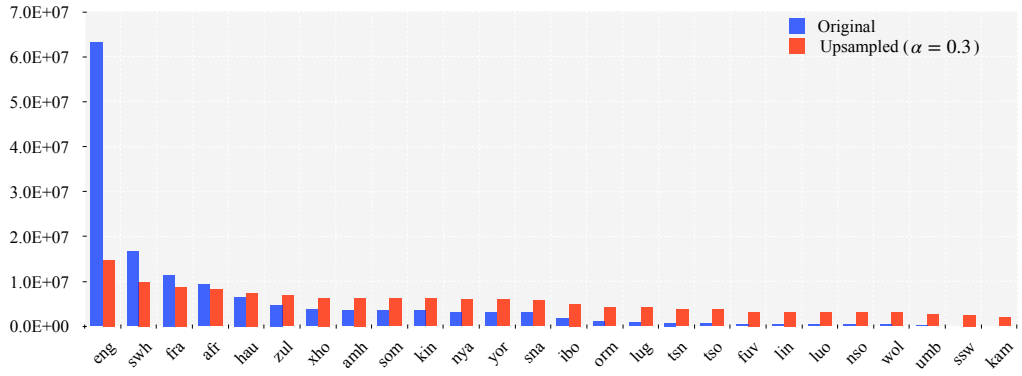

Figure 1: Number of sentences in each language and the upsampled distribution with the smoothing rate of $\alpha = 0.3$.

2.2 Data Preprocessing

We preprocess the raw and potentially noisy data by four steps, namely, reformatting, deduplication, language detection, and length limitation. Details are elaborated as below.

Reformatting. The raw data is stored in various alignment structures, including HTML, JSON, and special spacing. To reduce data noise, we reformat all data into a line-by-line tight structure and realign those missing paired ones.

Deduplication. We remove the duplicated sentences (pairs) in each monolingual and parallel dataset. This aims to reduce information redundancy so that the MNMT models can be trained more efficiently.

Language Detection. Previous studies suggest that incorrect languages in training data induce translation uncertainty for both bilingual (Ott et al., 2018) and multilingual (Wang et al., 2022) NMT models. Therefore, we conduct language detection for all the datasets using `langid`[4]. Since `langid` neither supports all African languages nor performs well when distinguishing two African languages, we adopt a simplified strategy: for the African datasets, we remove sentences (pairs) that are identified as languages other than the 24 designated African languages. In other words, sentences (pairs) in one African language identified as another African language are also considered valid. For English and French datasets, we strictly restrict the correct languages as themselves, i.e., English and French, respectively.

Length Limitation. After multilingual tokenization, we conduct further filtering and retain sentence pairs with tokens between 4 (Wu et al., 2019) and 512 (Yang et al., 2021), as well as the length ratio below 3.

2.3 Multilingual Tokenization

To tokenize the multilingual sentences, we follow (Conneau et al., 2020) to train a Sentence Piece Model (SPM) and apply it directly on the preprocessed text data for all languages. However, the distribution of data across languages is heavily long-tailed, as shown in Figure 1. To balance the vocabulary bandwidth between high-resource and low-resource languages, we follow Conneau et al. (2020) to upsample the low-resource languages with a smoothing rate of $\alpha = 0.3$ over the original distribution when training the SPM model. We use a shared vocabulary with 128K tokens for the 26 languages, and also append 32 special tokens (i.e., "TBD0" to "TBD31") for including extra tasks or data (e.g., tagged-BT (Caswell et al., 2019)).

3 Approach

3.1 Data Augmentation

We adopt data augmentation (DA) to address the first challenge, i.e., "**The absence of training data for some to-be-evaluated language pairs**".

Specifically, we use back-translation (BT) (Sennrich et al., 2016) and self-training (ST) (Jiao et al., 2020, 2021, 2022) to construct synthetic data. However, previous study by Caswell et al. (2019) suggests that the translationese issue in BT limits the performance, which can be mitigated with a special tag at source side (i.e., tagged-BT). To simplify the tagging procedure for the two opposite directions of each language pair, we use both BT and ST for each language pair (Wu et al., 2019) and append a special token at the synthetic side of sentence pairs. Formerly, for a language pair (S, T) with the bitext data $\{\mathbf{x}, \mathbf{y}\}$, the synthetic data by BT and ST will

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

be $\{[\mathbf{x}'; \langle \texttt{DA} \rangle], \mathbf{y}\}$ and $\{\mathbf{x}, [\mathbf{y}'; \langle \texttt{DA} \rangle]\}$, where $\langle \texttt{DA} \rangle$ denotes the special tag for data augmentation.

We conduct data augmentation for both English-centric and non English-centric language pairs. For English-centric language pairs, we randomly sample up to 1.0M English and non-English monolingual sentences from the training corpora for BT and ST, respectively. As for non English-centric language pairs, we translate the English side of English-centric pairs to non-English languages and construct up to 0.5M BT and ST sentence pairs, respectively. Generally, the augmented data is included in **Large-234** to train the MNMT models. However, the translation quality of those English-centric directions is also unreliable due to the limited data sizes, which may harm the performance of subsequent MNMT models. Besides, adding more synthetic data and language directions also slows down the convergence of the MNMT models. Instead, we use the resulting MNMT models as backbones to finetune on the clean datasets.

3.2 Distributionally Robust Optimization

We adopt the distributionally robust optimization (DRO) (Oren et al., 2019; Zhou et al., 2021) technique to address the second challenge, i.e., "**The uneven optimization of language pairs due to data imbalance** ".

Generally, temperature-based sampling (Arivazhagan et al., 2019; Conneau et al., 2020) is adopted to balance the training data across language pairs, which samples data from the smoothed data distribution as, $p_{\tau,i} = \frac{|D_i|^{1/\tau}}{\sum_j |D_j|^{1/\tau}}$. This is equivalent to optimizing the re-weighted objective:

$$\mathcal{L}_\tau(\theta; D_{\text{train}}) = \sum_{i \leq N} p_{\tau,i} \mathcal{L}(\theta; D_i), \qquad (1)$$

where $|D_i|$ is the training data size of the i-th language pair, and τ denotes the temperature rate. Obviously, $\tau = 1$ corresponds to the original data distribution while $\tau = \infty$ represents uniform sampling. In practice, $\tau > 1$ is adopted to oversample the low-resource language pairs, which significantly affects the results and needs to be tuned for different settings.

Even if we can build a completely balanced dataset across language pairs, the varied task difficulty and cross-lingual similarity determine that the language pairs will still be optimized unevenly. DRO can address such a problem. In contrast to temperature sampling which optimizes over a fixed training data distribution, DRO aims to find a model θ that can perform well on an entire set of potential test distributions, i.e., $\mathcal{U}(p^{\text{train}})$, which is usually called *uncertainty set*. We adopt DRO with the χ^2-uncertainty set introduced by Zhou et al. (2021), and reproduce the implementation for the practical many-to-many translation scenario.[5] Similarly, we also incorporate the baseline losses calculated from a pretrained MNMT model to stabilize the training process of DRO.

Table 2: Language family grouping.

Group	Target Languages
1	eng, fra
2	afr, nso, sna, ssw, tsn, tso, xho, zul
3	amh, luo, orm, som, swh, wol
4	fuv, hau, ibo, yor
5	kam, kin, lin, lug, nya, swh, umb

3.3 Language Family Grouping

We adopt language family grouping (LFG) to alleviate the third challenge, i.e., " **The curse of multilinguality**" (Conneau et al., 2020).

Specifically, we divide the target languages into 5 groups (see Table 2) based on Table 1. This is partially inspired by Eriguchi et al. (2022), which factorizes the many-to-many translation scenario (with $N \times N$ directions) into N many-to-one scenarios by training a translation model for each. Since we have 26 languages involved in this shared task, factorizing the many-to-many scenario by the family of target languages is a more efficient choice. Since **swh** appears in both HCEA and CA, we include it in both Group-2 and Group-5 for training models. During inference, our scripts will automatically select the model of corresponding group according to the target language to be evaluated. Note that **swh** is only routed to Group-2 in inference.

4 Experiments

4.1 Settings

Model. We adopt the standard sequence-to-sequence Transformer (Vaswani et al., 2017) as our architecture. For the **Base-146** scale, we use a deep encoder of 24 layers and a relatively shallow decoder of 12 layers (Yang et al., 2021), with an embedding size of 1024, the feed-forward network

[5]The referred study only supports one-to-many and many-to-one translation scenarios on very small multilingual translation datasets.

Table 3: Evaluation results of our models on the devtest in terms of BLEU and ChrF++.

Model	X-ENG	ENG-X	X-FRA	FRA-X	X-X	ALL
	22	22	4	4	48	100
TRANSF-DEEP	23.37/46.80	17.19/41.07	20.20/43.18	16.07/41.93	10.69/33.90	15.50/39.01
BORDERLINE-DEEP	25.87/48.83	18.24/42.05	22.31/44.91	16.74/42.32	11.66/34.89	16.86/40.23
BORDERLINE-DWIDE	28.11/51.30	19.02/42.92	24.74/47.58	17.22/43.23	12.03/34.94	17.82/41.13
BORDERLINE-DWIDE w/ LFG	28.26/51.37	19.38/43.37	24.87/47.74	17.48/43.72	12.04/35.01	**17.95/41.31**

size of 4096, and 16 attention heads (i.e., 0.59B parameters). To stabilize the training of deep models, we follow Wang et al. (2019) to use pre-layer-normalization (PLN) for both encoder and decoder layers. For the **Large-234** scale, we enlarge the embedding size to 1536 to support more language pairs, which results in 1.02B parameters. By default, we call these two models as TRANSF-DEEP and TRANSF-DWIDE. The final models developed by our approaches are renamed as BORDERLINE-DEEP and BORDERLINE-DWIDE for clarity.

Training. We train the MNMT models with the Adam optimizer (Kingma and Ba, 2014) ($\beta_1 = 0.9, \beta_2 = 0.98$). The learning rate is set as 1e-4 with a warm-up step of 4000, followed by inverse square root decay. The models are trained with a dropout rate of 0.1 and a label smoothing rate of 0.1. All experiments are conducted on 32 NVIDIA A100 GPUs. Since the bitext data (≈130M) for this year's shared task is less than 1/10 of that for last year's (≈1.7B), we decide a batch size to be about 1/10 of that used in (Yang et al., 2021). Specifically, we use 2048 max-tokens per GPUs and accumulate the gradients for every 8 steps to simulate the large batch size of 512K tokens. For language family grouping, we use the batch size of 131K tokens for each model. For translation models trained by empirical risk minimization (ERM) on the original training data, we upsample low-resource language pairs with the smoothing rate $\alpha = 0.3$ (Conneau et al., 2020). For those by DRO, we adopt the χ^2-uncertainty set with the distribution divergence bounded by $\rho = 0.1$. We use the ERM model to calculate the baseline losses for DRO. We train these two kinds of models for at least 100K updates, upon which we may finetune for additional updates.

Evaluation. We use the dev and devtest of Flores-200 benchmark[6] as our validation and test sets, and evaluate the MNMT models on the averaged last 10 checkpoints with sentencepiece BLEU and ChrF++. The sentencepiece model for evaluation also comes from the Flores-200 benchmark. The beam search process is performed with a beam size of 4 and a length penalty of 1.0. Similar as the official competition results, we report our results by average-to-eng (X-ENG), average-from-eng (ENG-X), average-to-fra (X-FRA), average-from-fra (FRA-X), average-african-to-african (X-X), and the average for ALL translation directions.

[6] https://github.com/facebookresearch/flores/tree/main/flores200

4.2 Results

We list the evaluation results of our final models on the devtest in Table 3. Both the baseline model TRANSF-DEEP and our BORDERLINE-DEEP model are trained for 200K updates, while the two BORDERLINE-DWIDE models are trained or finetuned for more than 300K updates.

Generally, our models outperform the baseline TRANSF-DEEP model significantly by up to +2.45 BLEU and +2.30 ChrF++ scores. By looking into each category, we have some interesting findings:

- By comparing BORDERLINE-DEEP and TRANSF-DEEP, we find that the improvement on X-ENG is much larger than that on ENG-X. Similar phenomenon is also observed for X-FRA and FRA-X. It suggests that while DRO can achieve even improvement for one-to-many or many-to-one scenarios (Zhou et al., 2021), it is heavily biased by the dominant languages (i.e., eng and fra) in the many-to-many scenario.

- By comparing BORDERLINE-DWIDE and BORDERLINE-DEEP, we find that enlarging the model capacity brings improvement to all categories but the most on X-ENG and X-FRA. It indicates that the *curse of multilinguality* cannot be well solved by simply increasing model capacity as the most benefits are still occupied by the dominant languages (i.e., eng and fra).

- Language family grouping (LFG) achieves more improvement on ENG-X and FRA-X than on the

Table 4: Official evaluation results of submissions on the blind test sets in terms of BLEU and ChrF++.

Submissions	X-ENG	ENG-X	X-FRA	FRA-X	X-X	ALL
#Lang-pairs	22	22	4	4	48	100
IIAI						
Primary	23.15/43.88	12.80/37.52	18.35/41.08	13.08/38.70	2.58/19.52	10.40/30.47
GMU						
Language	25.83/46.50	12.00/35.33	20.83/42.45	10.53/33.58	7.70/29.94	13.28/35.42
Family	25.88/46.55	11.98/35.30	20.73/42.30	10.75/34.03	7.68/29.92	13.28/35.42
Borderline (Ours)						
Contrastive	25.84/47.46	13.85/39.05	21.00/44.10	13.85/39.58	8.03/30.93	13.98/37.23
Primary	26.05/47.56	14.06/39.53	21.13/44.05	14.05/40.10	8.04/31.04	**14.09/37.42**

Table 5: Ablation study of our models with various strategies on the devtest. CT: continuous training; FT: finetuning; T-Enc: target language tags at encoder; LFG: language family grouping.

ID	Model	Step	BLEU	Δ
①	TRANSF-DEEP	100K	15.03	-/-
②	+ CT	100K	15.50	+0.47
③	+ FT on large-234	100K	14.65	-0.38
④	+ DRO	100K	16.71	+1.68
⑤	+ CT	100K	16.86	**+1.83**
⑥	+ T-Enc	100K	16.67	+1.64
⑦	TRANSF-DWIDE	100K	14.66	-/-
⑧	+ DRO	100K	15.81	+1.15
⑨	+ FT on base-146	200K	17.62	+2.96
⑩	+ FT on eval-106	50K	17.82	+3.16
⑪	+ LFG	-/-	17.95	+3.29

other categories, which confirms its effectiveness in alleviating the curse of multilingualty issue.

Ablation Study. We present detailed ablation studies to investigate the effectiveness of various strategies, not only the three introduced in §3 but also some tricks. The results are listed in Table 5, where the lines marked in blue (i.e., ②, ⑤, ⑩ and ⑪) correspond to the four models in Table 3. We list our observations as below:

- ③ *vs.* ②: Directly finetuning the TRANSF-DEEP model on the **Large-234** dataset induces the performance drop. One possible reason is that **Large-234** introduces much more translation directions, aggravating the *curse of multilinguality* issue. Another reason is the low-quality data by data augmentation (§3.1), which harms the optimization of models. Therefore, we only use **Large-234** to pretrain the TRANSF-DWIDE model and then finetune on the cleaner **Base-146** and **Eval-106** datasets.

- ⑥ *vs.* ⑤: Previous studies (Wang et al., 2022) suggest that attaching target language tags at encoder (i.e., T-Enc) benefits the zero-shot translation performance, indicating a stronger cross-lingual transfer ability. However, we do not see any improvement of our models with T-Enc. The reason could be that, traditional studies on many-to-many translations are mainly conducted on the datasets with only one central language while we are now handling multiple central languages, making it a more complex scenario.

- ⑨ *vs.* ⑩ *vs.* ⑪: Finetuning on **Eval-106** slightly outperforms that on **Base-146** and the performance can be further improved with language family grouping. Obviously, as we reduce the language pairs involved in a single model, the *curse of multilinguality* is alleviated.

Submissions. The BORDERLINE-DWIDE and BORDERLINE-DWIDE w/ LFG models shown in Table 3 (i.e., contrastive and primary versions) are submitted for official evaluation on the blind test sets. Table 4 summarizes the evaluation results of our submissions, where our models outperform the other teams' across all the evaluation groups. Finally, we achieve the **1st** place in this track.

5 Conclusion

In this paper, we describe Tencent's multilingual machine translation systems for the WMT22 shared task on Large-Scale Machine Translation Evaluation for African Languages. We address three key challenges of this task by data augmentation, distributionally robust optimization (DRO), and language family grouping, respectively, to develop our MNMT models. Our submissions won the **1st** place in the **constrained** track. Extensive analyses also point out the drawbacks of larger models and DRO in addressing the curse of multilinguality, which warrants further research in the future.

Acknowledgments

We sincerely thank the evaluation group for their hard work, including but not limited to the instant communication, environment configuration, and official evaluation. We also thank the anonymous reviewers for their insightful suggestions on various aspects of this report.

References

David Ifeoluwa Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta Costa-Jussá, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Safiyyah Saleem, and Holger Schwenk. 2022. Findings of the WMT 2022 shared task on large-scale machine translation evaluation for african languages. In *WMT*.

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*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *WMT*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Akiko Eriguchi, Shufang Xie, Tao Qin, and Hany Hassan Awadalla. 2022. Building multilingual machine translation systems that serve arbitrary xy translations. In *NAACL*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*

Wenxiang Jiao, Xing Wang, Shilin He, Irwin King, Michael Lyu, and Zhaopeng Tu. 2020. Data rejuvenation: Exploiting inactive training examples for neural machine translation. In *EMNLP*.

Wenxiang Jiao, Xing Wang, Shilin He, Zhaopeng Tu, Irwin King, and Michael R Lyu. 2022. Exploiting inactive examples for natural language generation with data rejuvenation. *IEEE/ACM TASLP*.

Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. 2021. Self-training sampling with monolingual data uncertainty for neural machine translation. In *ACL*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Z. Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv*.

Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust language modeling. In *EMNLP-IJCNLP*.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *ICML*.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised nmt using shared encoder and language-specific decoders. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *ACL*.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Xu Chen, Sneha Kudugunta, N. Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. *ACL*.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *FINDINGS*.

Nllb Team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*.

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 *NeurIPS*.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019. Learning deep transformer models for machine translation. In *ACL*.

Wenxuan Wang, Wenxiang Jiao, Shuo Wang, Zhaopeng Tu, and Michael R Lyu. 2022. Understanding and mitigating the uncertainty in zero-shot translation. *arXiv*.

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

Yiren Wang, ChengXiang Zhai, and Hany Hassan Awadalla. 2020b. Multi-task learning for multilingual neural machine translation. *EMNLP*.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. In *EMNLP-IJCNLP*.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, et al. 2021. Multilingual machine translation systems from microsoft for wmt21 shared task. In *WMT*.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *ICLR*.

Chunting Zhou, Daniel Levy, Xian Li, Marjan Ghazvininejad, and Graham Neubig. 2021. Distributionally robust multilingual machine translation. In *EMNLP*.