PARADISE: Exploiting Parallel Data for Multilingual Sequence-to-Sequence Pretraining

Machel Reid
The University of Tokyo
machelreid@weblab.t.u-tokyo.ac.jp

Mikel Artetxe
Meta AI
artetxe@fb.com

Abstract

Despite the success of multilingual sequence-to-sequence pretraining, most existing approaches rely on monolingual corpora, and do not make use of the strong cross-lingual signal contained in parallel data. In this paper, we present PARADISE (**PARA**llel & **D**enoising **I**ntegration in **SE**quence-to-sequence models), which extends the conventional denoising objective used to train these models by (i) replacing words in the noised sequence according to a multilingual dictionary, and (ii) predicting the reference translation according to a parallel corpus instead of recovering the original sequence. Our experiments on machine translation and cross-lingual natural language inference show an average improvement of 2.0 BLEU points and 6.7 accuracy points from integrating parallel data into pretraining, respectively, obtaining results that are competitive with several popular models at a fraction of their computational cost.[1]

1 Introduction

Multilingual sequence-to-sequence pretraining has achieved strong results both in cross-lingual classification (Xue et al., 2021) and machine translation (Liu et al., 2020). These models are usually pretrained on combined monolingual corpora in multiple languages using some form of denoising objective. More concretely, they noise each sequence x with a noising function g_ϕ, and maximize the probability of recovering x given $g_\phi(x)$:

$$\ell_{\text{mono}}(x) = -\log P\big(x|g_\phi(x)\big) \qquad (1)$$

Common noising functions include sentence-permutation and span masking (Lewis et al., 2020; Liu et al., 2020).

While these methods obtain strong cross-lingual performance without parallel data, they are usually trained at a scale that is prohibitive for most NLP practitioners. At the same time, it has been argued that the strict unsupervised scenario is not realistic (Artetxe et al., 2020), and parallel data could provide a stronger signal and make training more efficient.

Motivated by this, we propose PARADISE, a pretraining method for sequence-to-sequence models that exploits both word-level and sentence-level parallel data. The core idea of our approach is to augment the conventional denoising objective introduced above by (i) replacing words in the noised sequence according to a bilingual dictionary, and (ii) predicting the reference translation rather than the input sequence. Despite their simplicity, we find that both techniques bring substantial gains over conventional pretraining on monolingual data, as evaluated both in machine translation and zero-shot cross-lingual transfer. Our results are competitive with several popular models despite using only a fraction of the compute, providing strong support for the importance of the inclusion of parallel information in smaller-scale multilingual pretraining methods.

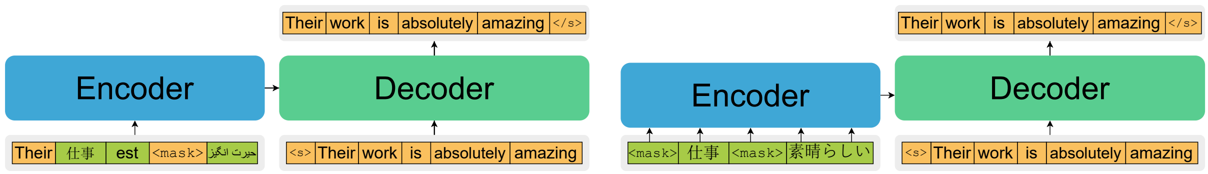
2 Proposed method

As illustrated in Figure 1, we propose two methods for introducing parallel data into pretraining: dictionary denoising and bitext denoising.

Dictionary denoising. Our first method encourages learning similar representations at the word-level by introducing anchor words through multilingual dictionaries (Conneau et al., 2020b). Let $D_l(w)$ denote the translation of word w into language $l \in L$ according to the dictionary D. Given the source sentence $x = (x_1, x_2, \ldots, x_n)$, we define its noised version $g_\psi(x) = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$, where $\tilde{x}_i = D_l(x_i)$ with probability $\frac{p_r}{|L|}$ and $\tilde{x}_i = x_i$ otherwise (i.e. we replace each word with its translation into a random language with probability

[1] Source code available at https://github.com/machelreid/paradise

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 800 - 810
July 10-15, 2022 ©2022 Association for Computational Linguistics



(a) Dictionary Denoising

(b) Bitext Denoising

Figure 1: Our proposed techniques for integrating parallel data into sequence-to-sequence pretraining.

p_r). We set $p_r = 0.4$. Given the dictionary-noised sentence, we train our model using the denoising auto-encoding objective in Eq. 1:

$$\ell_{\text{dict}}(x) = -\log P\big(x|g_\phi(g_\psi(x))\big) \tag{2}$$

Bitext denoising. Our second approach encourages learning from both monolingual and parallel data sources, by including translation data in the pretraining process. Given a source-target bitext pair (x, y) in the parallel corpus, assumed to be semantically equivalent, we model the following:

$$\ell_{\text{bitext}}(x, y) = -\log P\big(y|g_\phi(x)\big) \tag{3}$$

in which we optimize the likelihood of generating the target sentence y conditioned on the noised version of the source sentence, $g_\phi(x)$.[2]

Combined objective. Our final objective combines ℓ_{mono}, ℓ_{dict} and ℓ_{bitext}.[3] Given that our corpus contains languages with varying data sizes, we sample sentences using the exponential sampling technique from Conneau and Lample (2019). We use $\alpha_{\text{mono}} = 0.5$ to sample from the monolingual corpus, and $\alpha_{\text{bitext}} = 0.3$ to sample from the parallel corpus. To prevent over-exposure to English on the decoder side when sampling from the parallel corpus, we halve the probability of to-English directions and renormalize the probabilities. In addition, given that we have fewer amounts of parallel data (used for ℓ_{bitext}) than monolingual data (used for ℓ_{mono} and ℓ_{dict}), we sample between each task using $\alpha_{\text{task}} = 0.3$.

3 Experimental settings

We pretrain our models on 20 languages (English, French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili, Urdu, Japanese, Basque, Romanian, Sinhala and Nepalese), and evaluate them on machine translation and cross-lingual classification.

3.1 Pretraining

Data. We use Wikipedia as our monolingual corpus, and complement it with OSCAR (Ortiz Suárez et al., 2020), and CC100 (Conneau et al., 2020a) for low-resource languages. For a fair comparison with monolingually pretrained baselines, we use the same parallel data as in our downstream machine translation experiments (detailed in §3.2). In addition, we train a separate variant (detailed below) using additional parallel data from ParaCrawl (Esplà et al., 2019), UNPC (Ziemski et al., 2016), CCAligned (El-Kishky et al., 2020), and OpenSubtitles (Lison and Tiedemann, 2016).[4] We tokenize all data using SentencePiece (Kudo and Richardson, 2018) with a joint vocabulary of 125k subwords. We use bilingual dictionaries from FLoRes[5] (Guzmán et al., 2019) for Nepalese and Sinhala, and MUSE[6] (Lample et al., 2018) for the rest of languages. Refer to Appendix A for more details.

Models. We use the same architecture as BART-base (Lewis et al., 2020), totaling ~196M parameters, and train for 100k steps with a batch size of ~520k tokens. This takes around a day on 32 NVIDIA V100 16GB GPUs. As discussed before, we train two variants of our full model: PARADISE, which uses the same parallel data as the machine translation experiments, and PARADISE++, which uses additional parallel data. To better understand the contribution of each objective, we train two additional models without dictionary denoising, which we name **PARADISE (w/o dict.)** and **PARADISE++ (w/o dict.)**, as well as a model without bitext denoising, which we name **PARADISE++ (only dict.)**. Finally, we train a baseline system using the monolingual objective alone,

[2]To make our pretraining sequence length consistent with ℓ_{mono} and ℓ_{dict}, we concatentate randomly sampled sentence pairs from the same language pair to fit the maximum length.

[3]We use the same noising function g_ϕ used by Lewis et al. (2020) and Liu et al. (2020).

[4]We cap the size of each language pair to 2GB.

[5]https://github.com/facebookresearch/flores

[6]https://github.com/facebookresearch/MUSE

Languages	En-Vi		En-Tr		En-Ja		En-Ar		En-Ne		En-Ro		En-Si		En-Hi		En-Es		En-Fr	
Data Source	**IWSLT15**		**WMT17**		**IWSLT17**		**IWSLT17**		**FLoRes**		**WMT16**		**FLoRes**		**IITB**		**WMT13**		**WMT14**	
Size	133K		207K		223K		250K		564K		608K		647K		1.56M		15M		41M	
Direction	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→
Random init.	23.6	24.8	12.2	9.5	10.4	12.3	27.5	16.9	7.6	4.3	34.0	34.3	7.2	1.2	10.9	14.2	32.1	31.4	37.0	38.9
mBART (ours)	29.1	31.5	21.3	15.8	15.7	17.3	32.1	19.2	10.3	6.1	34.3	34.9	11.0	2.7	20.2	19.0	29.8	30.4	36.0	38.2
PARADISE	**30.0**	**32.6**	**23.5**	**17.2**	**17.2**	**19.2**	**35.3**	**21.1**	**13.7**	**7.9**	**35.9**	**36.5**	**14.0**	**3.7**	**23.6**	**20.7**	**32.6**	**32.7**	**37.8**	**39.8**

Table 1: Machine translation results. Random initialization numbers taken from Liu et al. (2020).

which we refer to as **mBART (ours)**. This follows the original mBART work (Liu et al., 2020), but is directly comparable to the rest of our models in terms of data and hyperparameters.

3.2 Downstream settings

Machine translation. Following Liu et al. (2020), we evaluate our models on sentence-level machine translation from and to English using the following datasets: IWSLT (Cettolo et al., 2015, 2017) for Vietnamese, Japanese and Arabic, WMT (Callison-Burch et al., 2009a,b; Bojar et al., 2016, 2017) for Spanish, French, Romanian and Turkish, FLoRes (Guzmán et al., 2019) for Sinhala and Nepalese, and IITB (Kunchukuttan et al., 2018) for Hindi. We report performance in BLEU as detailed in Appendix C.

Cross-lingual classification. We evaluate our models on zero-shot cross-lingual transfer on XNLI (Conneau et al., 2018) and PAWS-X[7] (Yang et al., 2019), where we finetune on English data and test performance on other languages. We develop a new approach for applying sequence-to-sequence models for classification: feeding the sequence into both the encoder and decoder, and taking the concatenation of the encoder's <s> representation and the decoder's </s> representation as the input of the classification head. We provide an empirical rationale for this in Appendix E. We finetune all models with a batch size of 64 and a learning rate of 2×10^{-5} for a maximum of 100k iterations, performing early stopping on the validation set.

4 Results

4.1 Machine translation

As shown in Table 1, PARADISE consistently outperforms our mBART baseline across all language pairs. Note that these two models have seen the exact same corpora, but mBART uses the parallel data for finetuning only, whereas PARADISE also uses it at the pretraining stage. This suggests that incorporating parallel data into pretraining helps learn better representations, which results in better downstream performance.

Lang. pair (En-XX)	Tr	Ro	Si	Hi	Es	Avg$_\Delta$
mBART (ours)	15.8	34.9	2.7	19.0	30.4	$20.6_{\pm 0.0}$
PARADISE (w/o dict.)	16.8	36.2	3.2	20.5	32.4	$21.8_{+1.2}$
PARADISE	17.2	36.5	3.7	20.7	32.7	$22.2_{+1.6}$
PARADISE++	19.0	37.3	4.2	20.7	33.0	$\mathbf{22.8}_{+2.2}$
Lang. pair (XX-En)	**Tr**	**Ro**	**Si**	**Hi**	**Es**	**Avg$_\Delta$**
mBART (ours)	21.3	34.3	11.0	20.2	29.8	$23.3_{\pm 0.0}$
PARADISE (w/o dict.)	23.2	35.6	13.2	22.3	31.6	$25.2_{+1.9}$
PARADISE	23.5	35.9	14.0	23.6	32.6	$25.9_{+2.6}$
PARADISE++	24.9	36.8	15.1	23.5	32.9	$\mathbf{26.6}_{+3.3}$

Table 2: Ablation results on machine translation.

Table 2 reports additional ablation results on a subset of languages. As can be seen, removing dictionary denoising hurts, but is still better than our mBART baseline. This shows that both of our proposed approaches—dictionary denoising and bitext denoising—are helpful and complementary. Finally, PARADISE++ improves over PARADISE, indicating that a more balanced corpus with more parallel data is helpful.

4.2 Cross-lingual classification

We report XNLI results in Table 3 and PAWS-X results in Appendix F. Our proposed approach outperforms mBART in all languages by a large margin. To our surprise, we also observe big gains in English. We conjecture that this could be explained by bitext denoising providing a stronger training signal from all tokens akin to ELECTRA (Clark et al., 2020), whereas monolingual denoising only gets effective signal from predicting the masked portion. In addition, given that we are using parallel data between English and other languages, PARADISE ends up seeing much more English text compared to mBART—yet a similar amount in the rest of languages—which could also contribute to

[7] Following Hu et al. (2021), we use English, German, Spanish, French and Chinese for PAWS-X.

Model	en	zh	es	de	ar	ur	ru	bg	el	fr	hi	sw	th	tr	vi	avg
mBART (ours)	77.5	68.0	70.7	68.8	66.7	62.2	68.6	72.1	69.6	70.1	63.4	62.6	66.6	65.0	69.7	68.1
PARADISE++ (only dict.)	79.6	70.9	74.9	75.4	64.4	66.0	69.0	72.2	75.4	73.4	63.9	65.1	70.9	69.0	72.1	70.8
PARADISE	**83.4**	73.8	77.6	76.0	**72.4**	65.1	74.0	74.4	73.2	**77.7**	**70.6**	66.2	70.4	72.1	75.3	73.5
PARADISE++ (w/o dict.)	83.3	72.9	77.2	75.7	64.4	**66.9**	73.4	74.8	75.7	**77.7**	68.5	67.4	71.0	73.3	75.0	73.1
PARADISE++	83.0	**74.0**	**79.0**	**76.5**	68.5	66.8	**74.3**	**76.0**	**76.4**	**77.7**	70.2	**70.5**	**72.3**	**74.2**	**75.4**	**74.3**

Table 3: Accuracy of zero-shot crosslingual classification on the XNLI dataset.

Model	**#Langs**	**Task**	**Params.**	**Est. GPU Days**	**Data (GB)**	**XNLI**	**PAWS-X**	**MT**
mBERT (Devlin et al., 2019)[†]	104	MLM	179M (0.9x)	—	60	65.4	86.2	—
MMTE (Siddhant et al., 2019)[†]	102	Translation	375M (1.9x)	—	5000	67.4	85.6	—
mT5-small (Xue et al., 2021)	101	Eq. 1	300M (1.5x)	—	27000	67.5	85.8	—
mT6 (Chi et al., 2021a)	94	SC+PNAT+TSC	300M (1.5x)	40 (1.3x)	2120	64.7	86.6	—
AMBER (Hu et al., 2021)	104	MLM+TLM	179M (0.9x)	1000 (31x)	100	71.6	89.2	—
XLM-15 (Conneau and Lample, 2019)[‡]	15	MLM+TLM	250M (1.3x)	450 (14x)	100	72.6	88.0	—
XLM-R-base (Conneau et al., 2020a)[‡]	100	MLM	270M (1.4x)	13K (406x)	2400	73.4	87.4	—
mBART (Liu et al., 2020)	25	Eq. 1	680M (3.5x)	4.5K (140x)	2400	—	—	23.5
mBART (ours)	20	Eq. 1	196M (1.0x)	32 (1.0x)	72	68.1	85.4	21.1
PARADISE	20	Eq. 1, 2, 3	196M (1.0x)	32 (1.0x)	81	73.5	89.0	23.1
PARADISE++	20	Eq. 1, 2, 3	196M (1.0x)	32 (1.0x)	95	**74.3**	**89.2**	**23.8**

Table 4: Comparison with prior work. † denotes results taken from Hu et al. (2020). ‡ denotes results taken from Hu et al. (2021). 1 GPU day = 1 day on an NVIDIA V100 GPU.

its better performance in this language. Finally, we observe that all of our different variants perform similarly in English, but incorporating dictionary denoising and using additional parallel data both reduce the cross-lingual transfer gap.

4.3 Comparison with prior work

So as to put our results into perspective, we compare our models with several popular systems from the literature. As shown in Table 4, our proposed approach obtains competitive results despite being trained at a much smaller scale. Just in line with our previous results, this suggests that incorporating parallel data makes pretraining more efficient given that we outperform XLM-R base, mT5, and mBART despite using less data/compute/model size. Interestingly, our method also outperforms XLM-15, MMTE, and mT6 which also use parallel data, as well as AMBER, showing evidence contrary to Hu et al. (2021)'s suggestion that using dictionaries may hurt performance. Detailed per-language results for each task can be found in Appendix F.

5 Related work

Most prior work on multilingual pretraining uses monolingual data only (Pires et al., 2019; Conneau et al., 2020a; Song et al., 2019; Liu et al., 2020; Xue et al., 2021). There have been several proposals to incorporate parallel data into encoder-only models (Lample and Conneau, 2019; Huang et al., 2019; Hu et al., 2021; Chi et al., 2021b), with some approaches replacing words according to a bilingual dictionary, similar to our dictionary denoising objective (Conneau et al., 2020b; Chaudhary et al., 2020; Dufter and Schütze, 2020). In contrast, we focus on sequence-to-sequence models, which we believe are more flexible and provide a more natural way of integrating parallel data. In that spirit, Siddhant et al. (2019) showed that vanilla machine translation models are already competitive in cross-lingual classification. Closer to our work, Chi et al. (2021a) incorporated parallel corpora into sequence-to-sequence pretraining by feeding concatenated parallel sentences to the encoder and using different masking strategies. In contrast, our approach feeds a noised sentence into the encoder, and tries to recover its translation in the decoder side, obtaining better results with a similar computational budget. Concurrent to our work, Kale et al. (2021) extended T5 to incorporate parallel corpora using a similar approach to our bitext denoising.

6 Conclusions

In this work, we proposed PARADISE, which introduces two new objectives to integrate parallel data into sequence-to-sequence pretraining. Experimental results on machine translation and cross-lingual classification show that PARADISE provides significant improvements over mBART-style pretraining on monolingual corpora, obtaining results that are competitive with several popular models at a much smaller scale. Given these findings, we encourage use of parallel data in smaller-scale multilingual pretraining work. In the future, we look to see if our improvements also hold at a larger scale.

Acknowledgements

We thank Junjie Hu, Jungo Kasai, and Victor Zhong for useful suggestions and comments. MR is grateful to the Masason Foundation for their support

References

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A Call for More Rigor in Unsupervised Cross-lingual Learning. *ArXiv*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009a. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009b. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

M. Cettolo, Marcello Federico, L. Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign.

M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and Marcello Federico. 2015. The iwslt 2015 evaluation campaign.

Aditi Chaudhary, Karthik Raman, Krishna Srinivasan, and Jiecao Chen. 2020. Dict-mlm: Improved multilingual pre-training using bilingual dictionaries.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021a. mt6: Multilingual pretrained text-to-text transformer with translation pairs. *arXiv preprint arXiv:2104.08692*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021b. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv:2003.10555 [cs]*.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

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*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLoRes Evaluation Datasets for Low-Resource Machine Translation: Nepali-English and Sinhala-English. *arXiv:1902.01382 [cs]*.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit Alignment Objectives for Multilingual Bidirectional Encoders. *arXiv:2010.07972 [cs]*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multitask Benchmark for Evaluating Cross-lingual Generalization. *arXiv:2003.11080 [cs]*.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Mihir Kale, Aditya Siddhant, Noah Constant, Melvin Johnson, Rami Al-Rfou, and Linting Xue. 2021. nmt5 – is parallel data still relevant for pre-training massively multilingual language models? *arXiv preprint arXiv:2106.02171*.

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.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. *arXiv:1901.07291 [cs]*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Arivazhagan, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2019. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. *arXiv:1905.02450 [cs]*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Data

We list data sources used for pretraining PARADISE++ in Table 5 (monolingual data) and Table 6 (parallel data).

B Pretraining hyperparameters

We use the Adam optimizer ($\epsilon = 10^{-6}, \beta = (0.9, 0.98)$), and warm up the learning rate to a peak of 7×10^{-4} after 10K iterations and then proceed to decay the learning rate with the polynomial decay schedule up until 100K iterations. All code and experiments are performed with `fairseq` (Ott et al., 2019). Following Liu et al. (2020), we add an additional layer-normalization layer on top of both the encoder and decoder to stabilize training with FP16 precision (Micikevicius et al., 2018). All models are trained on 32 V100 16GB GPUs and take 24 hours to finish training.

C Machine translation evaluation

Following Liu et al. (2020), we use detokenized SacreBLEU (Post, 2018) for all languages unless specified otherwise next. For Japanese we use KyTea[8], for Nepalese, Sinhala, and Hindi we use Indic-NLP[9], for Arabic we use the QCRI Arabic Normalizer[10,11], and for Romanian we use Moses tokenization and script normalization following Sennrich et al. (2016); Liu et al. (2020).

D Machine translation finetuning

We finetune our models using the same setup as mBART, warming up the learning rate to 3×10^{-5} over 2500 iterations and then decaying with a polynomial schedule. We use 0.3 dropout and label smoothing $\epsilon = 0.2$.

E Comparison of finetuning approaches

Table 7 compares our proposed finetuning approach, which combines the representations from both the encoder and the decoder (see §3), to using either of them alone.[12] While prior work either minimally used the decoder if at all (Siddhant et al., 2019; Xue et al., 2021), or only added a classification head on top of the decoder (Lewis et al., 2020), we find that combining them both works best.

[8] http://www.phontron.com/kytea/

[9] https://github.com/anoopkunchukuttan/indic_nlp_library

[10] https://github.com/qntfy/gomosesgo

[11] https://alt.qcri.org/tools/arabic-normalizer/

[12] For *decoder-only*, we feed the input sequence to both the encoder and the decoder, but add a classification head on top of the decoder only, following Lewis et al. (2020).

Language	Data source	Data size (GB)
En	Wiki	14G
De	Wiki	5.9G
Fr	Wiki	4.5G
Es	Wiki	3.7G
Ja	Wiki	3.0G
Ru	Wiki	6.2G
Ar	Wiki	1.7G
Ne	CC100	3.8G
Si	CC100	3.7G
Ro	Wiki+WLM	2.5G
Zh	Wiki+WLM	4.4G
El	Wiki+WLM	2.9G
Eu	Wiki+OSCAR	0.6G
Bg	Wiki+OSCAR	2.5G
Hi	Wiki+OSCAR	2.3G
Sw	Wiki+CC100	1.1G
Th	Wiki+OSCAR	2.4G
Ur	Wiki+OSCAR	1.9G
Vi	Wiki+OSCAR	2.8G
Tr	Wiki+OSCAR	2.4G
Total	—	72G

Table 5: Monolingual Data Statistics. Wiki refers to Wikipedia, and WLM refers to the News Crawl data from CommonCrawl used in WMT.

Language	Data source	Data size (GB)	# Pairs
Ar	UNPC	2.0G	5554595
Bg	ParaCrawl	1.9G	6470710
De	ParaCrawl	2.0G	9685483
El	ParaCrawl	2.0G	6676200
Es	ParaCrawl	2.0G	9138031
Eu	OPUS	0.1G	585210
Fr	ParaCrawl	2.0G	8485669
Hi	IITB	0.4G	1609682
Ja	JParaCrawl	2.0G	6366802
Ne	CCAligned	0.2G	487157
Ro	ParaCrawl	1.3G	6160525
Ru	ParaCrawl	1.6G	5377911
Si	CCAligned	0.2G	619730
Sw	OPUS	0.2G	699719
Th	OpenSubtitles	0.4G	3281533
Tr	OpenSubtitles	2.0G	32077240
Ur	CCAligned	0.3G	1371930
Vi	OpenSubtitles	0.2G	3505276
Zh	UNPC	2.0G	7706183
Total	—	23G	126882448

Table 6: Parallel Data Statistics

Model	avg	Δ
PARADISE++ (*encoder-decoder*)	74.3	—
decoder-only	73.8	-0.5
encoder-only	72.0	-2.3

Table 7: Ablation of finetuning methods on XNLI.

F Additional results

We list detailed results by language in this section with results on XNLI in Table 8, PAWS-X in Table 9, and our machine translation ablation (with mBART (Liu et al., 2020) results included) in Table 10. We note that mBART underperforms XLM-R-large on XNLI, however that may be attributed to the fact that XLM-R was trained for much longer rather than the architectural design.

Models	en	zh	es	de	ar	ur	ru	bg	el	fr	hi	sw	th	tr	vi	avg
mBERT	80.8	67.8	73.5	70.0	64.3	57.2	67.8	68.0	65.3	73.4	58.9	49.7	54.1	60.9	69.3	65.4
MMTE	79.6	69.2	71.6	68.2	64.9	60.0	66.2	70.4	67.3	69.5	63.5	61.9	66.2	63.6	69.7	67.5
mT5-small	79.6	65.8	72.7	69.2	65.2	59.9	70.1	71.3	68.6	70.7	62.5	59.7	66.3	64.4	66.3	67.5
AMBER	84.7	71.6	76.9	74.2	70.2	61.0	73.3	74.3	72.5	76.6	66.2	59.9	65.7	73.2	73.4	71.6
XLM-15 (MLM+TLM)	84.1	68.8	77.8	75.7	70.4	62.2	75.0	75.7	73.3	78.0	67.3	67.5	70.5	70.0	73.0	72.6
XLM-100	82.8	70.2	75.5	72.7	66.0	59.8	69.9	71.9	70.4	74.3	62.5	58.1	65.5	66.4	70.7	69.1
XLM-R-base	83.9	73.6	78.3	75.2	71.9	65.4	75.1	76.7	75.4	77.4	69.1	62.2	72.0	70.9	74.0	73.4
mBART	87.7	76.4	81.5	79.8	75.5	—	78.9	—	—	80.6	73.0	—	—	76.1	77.4	—
XLM-R-large	**88.7**	**78.2**	**83.7**	**82.5**	**77.2**	**71.7**	**79.1**	**83.0**	**80.8**	**82.2**	**75.6**	**71.2**	**77.4**	**78.0**	**79.3**	**79.2**
mBART (ours)	77.5	68.0	70.7	68.8	66.7	62.2	68.6	72.1	69.6	70.1	63.4	62.6	66.6	65.0	69.7	68.1
PARADISE (w/o dict.)	**83.3**	72.9	77.2	75.7	64.4	**66.9**	73.4	74.8	75.7	**77.7**	68.5	67.4	71.0	73.3	75.0	73.1
PARADISE	83.0	**74.0**	**79.0**	**76.5**	**68.5**	66.8	**74.3**	**76.0**	**76.4**	**77.7**	**70.2**	**70.5**	**72.3**	**74.2**	**75.4**	**74.3**

Table 8: Accuracy of zero-shot crosslingual classification on the XNLI dataset. Bold numbers highlight the highest scores across languages on the existing models (upper part) and PARADISE variants (bottom part). Results for previous work are sourced from Hu et al. (2020, 2021); Xue et al. (2021).

Model	de	en	es	fr	zh	Avg
mBERT	85.7	94.0	87.4	87.0	77.0	86.2
MMTE	85.1	93.1	87.2	86.9	75.9	85.6
mT5-small	86.2	92.2	86.1	86.6	77.9	85.8
AMBER	89.4	**95.6**	89.2	**90.7**	80.9	89.2
XLM-15	88.5	94.7	89.3	89.6	78.1	88.0
XLM-100	85.9	94.0	88.3	87.4	76.5	86.4
XLM-R-base	87.0	94.2	88.6	88.7	78.5	87.4
XLM-R-large	**89.7**	94.7	**90.1**	90.4	**82.3**	**89.4**
PARADISE++	89.1	94.3	89.6	90.6	**82.3**	89.2

Table 9: Accuracy of zero-shot cross-lingual classification on PAWS-X. Bold numbers highlight the highest scores across languages on the existing models (upper part) and PARADISE variants (bottom part). We source baseline results from Hu et al. (2020, 2021); Xue et al. (2021).

Lang. Pair	En-Tr	En-Ro	En-Si	En-Hi	En-Es	Tr-En	Ro-En	Si-En	Hi-En
mBART (ours)	15.8	34.9	2.7	19.0	30.4	21.3	34.3	11.0	20.2
PARADISE (w/o dict.)	16.8	36.2	3.2	20.5	32.4	23.2	35.6	13.2	22.3
PARADISE	17.2	36.5	3.7	20.7	32.7	23.5	35.9	14.0	**23.6**
PARADISE++	**19.0**	37.3	**4.2**	20.7	33.0	**24.9**	36.8	**15.1**	23.5
mBART	17.8	**37.7**	3.3	**20.8**	**34.0**	22.5	**37.8**	13.7	23.5

Table 10: Ablation results on machine translation. Note that mBART is trained with 140x more compute and 3.5x more parameters.