LAFT: Cross-lingual Transfer for Text Generation by Language-Agnostic Finetuning

Xianze Wu[1*] **Zaixiang Zheng**[2] **Hao Zhou**[3†] and **Yong Yu**[1‡]
[1]Shanghai Jiao Tong University [2]Bytedance AI Lab
[3]Insititute for AI Industry Research, Tsinghua University
{xzwu,yyu}@apex.sjtu.edu.cn
zhengzaixiang@bytedance.com
zhouhao@air.tsinghua.edu.cn

Abstract

Multilingual language pretraining enables possibilities of transferring task knowledge learned from a rich-resource source language to the other, particularly favoring those low-resource languages with few or no task annotated data. However, knowledge about language and tasks encoded is strongly entangled in multilingual neural representations, thereby the learned task knowledge falsely correlated to the source language, falling short of cross-lingual transferability. In this paper, we present a novel *language-agnostic finetuning* (LAFT) to facilitate zero-resource cross-lingual transfer for text generation. LAFT performs *language-agnostic task acquisition* to isolate task learning completely from the source language, and then *language specification* for better generation for specified languages. Experiments demonstrate that the proposed approach facilitates a better and parameter-efficient transferability on two text generation tasks.

1 Introduction

Deep learning has boosted the development of natural language generation (NLG), giving rise to its applications to a broad range of tasks (Brown et al., 2020; Liu et al., 2020; Xue et al., 2021), e.g., summarizing a lengthy news article. Annotated data is essential for learning neural NLG models. However, the vast bulk of available data is normally presented in English, making data scarcity in other languages a significant difficulty. Therefore, cross-lingual transfer, the ability to transfer knowledge learned in a rich-resource source language (typically English) to other, unseen target languages, has enormous practical significance.

The recent success of multi-lingual pre-trained language models (MPLMs) (Liu et al., 2020; Conneau et al., 2020; Xue et al., 2021) enables possibilities for such zero-resource cross-lingual transfer in a "pretrainig-finetuning" paradigm. Specifically, thanks to that MPLMs can learn plausible multilingual representations for any languages involved in multi-lingual pretraining, finetuning a MPLM on task annotated data in English can exhibit immediate task performance on other languages. However, despite its appealing results on natural language understanding, the transferring performance remains unsatisfactory on language generation tasks.

The neural NLG pipeline consists of three sequential steps: a) understanding input text (e.g., a news article), b) manipulating semantics in accordance with the task (e.g., filtering out redundant content while retaining content of the main idea), and c) generating text result (e.g., abstractive summary). As a result, we suggest that learning a generation task essentially bolts down to learning how to manipulate the input semantic for the following generation. However, due to the highly entangled nature of semantic information and language information learned in multilingual representations, knowledge of a downstream task learned by finetuning would inevitably be correlated to the source language, thus harming the ability to transfer to unseen target languages.

In this paper, we propose the *language-agnostic finetuning* (LAFT). The key idea is to completely isolate acquiring task knowledge for an MPLM from the source language, and then add the language information back for generation. Given a text generation task and its annotated data in the source language, LAFT consists of two stages:

- **Language-agnostic task acquisition.** An extra task module is added to the MPLM. The module learns to manipulate semantic content given the task without considering any information about the source language.
- **Language specialization.** We then incorporate language information back into the

[*] Work was done during Xianze's internship at ByteDance AI Lab, mentored by Zaixiang.

[†] Work was done while at ByteDance.

[‡] Corresponding authors.

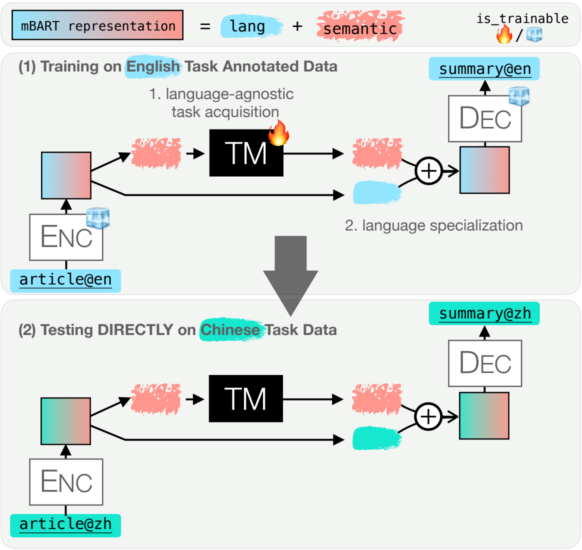
mBART representation
=
lang
+
semantic
is_trainable
/
(1) Training on English Task Annotated Data
1. language-agnostic task acquisition
TM
summary@en
DEC
ENC
article@en
2. language specialization
(2) Testing DIRECTLY on Chinese Task Data
TM
summary@zh
DEC
ENC
article@zh

Figure 1: Illustration of LAFT for mBART. (1) Training on source language task annotated data. (2) Trained model can be directly evaluated for target language.

task module's language-agnostic representation, helping the decoder to better generate the resulting content in the specified language.

We evaluate our zero-resource cross-lingual transfer approach in two scenarios: zero-shot and translate-train, which differ in terms of the existence of machine translation systems. Experimental results show that the proposed method facilitates a better and parameter-efficient transferability on abstractive summarization (+up to 0.71 ROUGE-L) and question generation (+up to 2.45 ROUGE-L), which could motivate further research that cross-lingual transfer necessitates careful consideration of task acquisition and language specialization,

2 Related Work

Most previous cross-lingual transfer research has succeeded on NLU rather than NLG. For both NLU and NLG, one solution is data augmentation that leverages data from the source language to the target language using translation systems or code-switching (Singh et al., 2019; Bornea et al., 2021; Qin et al., 2020). Some NLU research aims to learn language-agnostic features that minimize the distance among features from different languages, by adversarial training (Keung et al., 2019; Chen et al., 2019), removing the language identity from the original multi-lingual representations (Libovický et al., 2020; Zhao et al., 2021; Yang et al., 2021; Tiyajamorn et al., 2021) or contrastive learning(Yu and Joty, 2021).

For NLG, one of the most promising findings of cross-lingual transfer is that multilingual machine translation systems trained on massive amount of multilingual data manifest emergent ability of unsupervised (Üstün et al., 2021) or zero-shot translation for those unseen language pairs (Gu et al., 2019; Chen et al., 2022). Such observations encourage researchers to design effective pretraining objectives favoring cross-lingual transfer for monolingual text generation tasks (e.g., summarization) (Chi et al., 2020; Lewis et al., 2020; Maurya et al., 2021), whereas the finetuning process receives little attention. Despite learning language-agnostic features for finetuning as in NLU is promising, language information, in contrast to NLU, is critical for NLG. If only language-agnostic features are used, the model will not be able to generate text in the specified language.

3 Methodology: LAFT

Figure 1 shows the overall workflow of LAFT when applying to mBART (Liu et al., 2020).[1] As illustrated, we first introduce an extra task module (TM), parameterized by two Transformer layers(Vaswani et al., 2017), between the encoder and decoder for **language-agnostic task acquisition** (§3.1), where the TM is expected to learn how to manipulate input semantic content given the task. We then perform **language specialization** by adding language information to the language-agnostic representation obtained by the TM, allowing the decoder to synthesize the resulting text in the provided language (§3.2).

3.1 Language-agnostic Task Acquisition

Our approach is inspired by Yang et al. (2021) that for an MPLM, the representations from the same language L tend to cluster together, which implies that they share vector space components that correspond to the language identity of the language L. This finding intuitively enables disentangling the semantic contents from language identity by removing the language components from the representation, which can be conducted as following two steps:

(1) Estimation of language component. Given a pretrained mBART, its encoder can be seen as a multi-lingual embedding system E. For

[1] In this paper, we primarily study the proposed language-agnostic finetuning on mBART, but the method can be applied to any encoder-decoder MPLMs.

each language L, we construct a language matrix $M_L \in \mathbb{R}^{n\times d}$ based on a collection of monolingual texts $\{t_L^i\}_{i=0}^n$, where the ith row of M_L is the sentence representation of t_L^i given by E. We then apply singular value decomposition (SVD) $M_L = U_L \Sigma_L V_L^T$, and extract the first k right singular vectors (i.e., columns of $V_L \in \mathbb{R}^{d\times d}$) as the shared components for language identity of L, denoted as $c_L \in \mathbb{R}^{d\times k}$.

(2) Removal of language component. Given a text $x_L = \{x_L^i\}$ from the language L, where x_L^i is the ith token of x_L, we denote the representation of x_L^i given by the encoder as e_L^i. The sentence representation e_L is obtained via the mean-pooling of $\{e_L^i\}$. Then we subtract the projection of e_L onto c_L from e_L^i as

$$r_L^i = e_L^i - c_L \frac{c_L^T e_L}{\|e_L\|_2}.$$

As a result, $\boldsymbol{r}_L = \{r_L^i\}$ is the language-agnostic representation as expected, which is then fed into the TM for learning the task:

$$\boldsymbol{h}_L = \mathrm{TM}(\boldsymbol{r}_L)$$

3.2 Language Specialization for Generation

The proposed language-agnostic task acquisition eases the transfer of task knowledge across language. Unlike NLU tasks, which can rely solely on semantic information for classification, language information is critical for NLG tasks since we want to generate text in a specific language. Thus, beside language-agnostic task acquisition, we also need to improve the model regarding its language generation ability. We refer to this as language specialization, which includes two aspects: (1) we integrate the subtracted language components into the TM's language-agnostic output, (2) we enhance the decoder with an extra language adapter.

Fusing with subtracted language components. We apply a fusion mechanism to add subtracted language components c_L back to the TM's output:

$$\mathbf{B}(h_L^i, c_L) = \mathbf{U}\left(\mathrm{ReLU}\left(\mathbf{D}([h_L^i, c_L])\right)\right) + h_L^i,$$

where $\mathbf{D} \in \mathbb{R}^{2d_h\times d_a}$ and $\mathbf{U} \in \mathbb{R}^{d_a\times d_h}$ are parametrized by two feed-forward layers. $\mathbf{B}(h_L^i, c_L)$ is then fed into the decoder.

Enhancing decoder with language adapter. The decoder is responsible for generating text in a given language. To promote the decoder to adapt to the fused representations, we incorporate a feed-forward layer based language adapter to each decoder layer (Pfeiffer et al., 2020a), which is jointly trained with the fusion mechanism.

3.3 Learning

Learning of LAFT contains two stages.

(1) *Unsupervised generation pretraining*. In this stage, we only allow the TM and fusion mechanism trainable while keeping the remainder of the model parameters frozen. We leverage *unsupervised data* from the source and target language. Following (Liu et al., 2020), we use a cross-entropy loss between the original document and the decoder's output given the corrupted document as input, which is constructed by applying "text infilling" noise to the original document (Liu et al., 2020).

(2) *Task finetuning*. In this stage, given *source language annotated task data*, we freeze the fusion mechanism and optimize the TM using the cross-entropy loss between the decoder's output and the ground-truth reference.

4 Experiments

We experiment on two NLG tasks, i.e., abstractive text summarization and question generation to evaluate our LAFT for cross-lingual transfer.

Datasets. For text summarization, we perform experiments on the XGIGA datasets. We choose its English part as the training set and its French and Chinese parts as the evaluation set. For question generation, we choose the XQG dataset (Chi et al., 2020). The XQG dataset consists of the English part and the Chinese part. We train models on English part and evaluate models on Chinese part.

We learn language specialization using `cc100` dataset (Conneau et al., 2020), from which we select a subset containing 1,000,000 sentences for Chinese, English and French respectively.

Baselines. We compare LAFT with the following baselines:

- mBART (full): directly finetuning the full parameters of mBART on English annotated data;
- mBART (enc): only finetuning the encoder parameters of mBART;
- TM + adv: using adversarial training instead of LAFT to force the output of TM to be language-agnostic.

More details are presented in Appendix.

Setting	Zero-shot		Trans-train	
Language	zh→zh	fr→fr	zh→zh	fr→fr
Baselines				
mBART (full)	43.82	33.40	47.33	42.8
mBART (enc)	45.85	36.55	47.09	42.11
TM + adv	31.41	36.71	**48.04**	43.04
LAFT	**46.37**	**40.78**	47.66	**43.10**

Table 1: Results of abstractive summarization. “full“: finetuning full model. “enc”: finetuning only encoder

Setting	Zero-shot	Trans-train
Language	zh→zh	zh→zh
Baselines		
mBART (full)	21.62	36.58
mBART (enc)	32.08	33.57
TM + adv	21.98	37.02
LAFT	**34.53**	**37.02**

Table 2: Results of question generation.“full“: finetuning full model. “enc”: finetuning only encoder.

Results of Zero-shot Setting. First, we evaluate models on the zero-shot cross-lingual transfer. Results of abstractive summarization and question generation are presented in Table 1 and Table 2, respectively. When a full mBART is fine-tuned, it runs the danger of incorrectly associating the task to the source language, resulting in poor transfer performance. Only Finetuning the encoder can somehow alleviate but does not fundamentally address the problem. LAFT, on the other hand, can learn task ability avoiding associating to the source language, which improves transferability for generation and outperforms baseline systems. Surprisingly, while the adversarial method is known to be good at removing language information, it fails miserably in the zero-shot case due to a lack of task data for each language, causing the model to degenerate into copying the input sequence regardless of languages.

Results of Translate-train Setting. We evaluate models on the translate-train setting to see if data augmentation by machine translation could further help. As shown in Table 1 and Table 2, we can observe that data augmentation can generally improve all approaches. Note that because pseudo task data for target languages is accessible in this setting, the adversarial method can function normally. Nevertheless, our LAFT still achieves comparable results with the adversarial method, demonstrating the effectiveness of the proposed method.

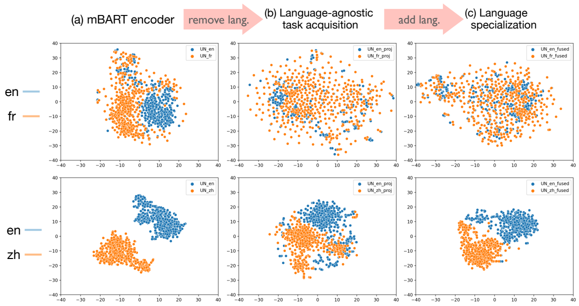

Figure 2: t-SNE (Van der Maaten and Hinton, 2008) visualization of representations.

Model	R-L (↑)	$\|\theta_{\text{trainable}}\|\%$ (↓)
mBART (full)	43.82	100%
mBART (enc)	45.85	19.2%
mBART (enc top-2)	44.85	3.8%
Adapter (Pfeiffer et al., 2020a)	43.05	4.3%
LAFT	**46.37**	**3.8%**

Table 3: Number of trained parameters and results on abstractive summarization. “enc top-2”: only finetuning the top two layers of the encoder.

Visualization of LAFT. To ensure that LAFT can yield language-agnostic representations, we visualize the representations before and after applying LAFT in Figure 2. As we can see, the original mBART encoder representation is distributed separately in terms of languages (Figure 2(a)). After removing language identity, the distribution of representations from different languages becomes closer, allowing the model to produce language-agnostic representations for task acquisition (Figure 2(b)). Finally, once language specialization is performed, the representations become language-aware thus distribute separately again, making it easier for the decoder to generate text in a specific language (Figure 2(c)).

Analysis of Parameter Efficiency. To demonstrate parameter efficiency of LAFT, we compare the performance of abstractive summarization with the number of training parameters. As shown in Table 3, our method yields the best ROUGE-L score with the fewest training parameters, demonstrating that LAFT results in a parameter-efficient model.

5 Conclusion

This paper proposes language-agnostic finetuning (LAFT) to facilitate zero-resource cross-lingual transfer for text generation. We finetune a task module only through the semantic contents of a multi-lingual representation. To achieve it, we utilize a disentangled-based and an adversarial-based method. Then we combine the information of a language with the task module's language-agnostic representation, allowing the model to generate text in the language. Experimental results show that language-agnostic finetuning results in a better and parameter-efficient transferability on two text generation tasks. The major limitation of our work is we only explore two target languages. We leave other languages for future work.

6 Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments.

References

Mihaela A. Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. 2021. Multilingual transfer learning for QA using translation as data augmentation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12583–12591. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2022. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 142–157. Association for Computational Linguistics.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3098–3112. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7570–7577. AAAI Press.

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

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1258–1268. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 1355–1360. Association for Computational Linguistics.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida I. Wang, and Luke Zettlemoyer. 2020. Pre-training via paraphrasing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jindrich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1663–1674. Association for Computational Linguistics.

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. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. 2021. Zmbart: An unsupervised cross-lingual transfer framework for language generation. In *Findings of the Association for Computational Linguistics:*

ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2804–2818. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 46–54. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: an adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7654–7673. Association for Computational Linguistics.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3853–3860. ijcai.org.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. XLDA: cross-lingual data augmentation for natural language inference and question answering. *CoRR*, abs/1905.11471.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774.

Ahmet Üstün, Alexandre Berard, Laurent Besacier, and Matthias Gallé. 2021. Multilingual unsupervised neural machine translation with denoising adapters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6650–6662. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 2021. A simple and effective method to eliminate the self language bias in multilingual representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5825–5832. Association for Computational Linguistics.

Tao Yu and Shafiq R. Joty. 2021. Effective fine-tuning methods for cross-lingual adaptation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8492–8501. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, *SEM 2021, Online, August 5-6, 2021*, pages 229–240. Association for Computational Linguistics.

A Appendix

Implement Details. We choose the mBART$_{\text{large}}$ model as the backbone model. The task module consists of two transformer layers, whose setting is the same as the transformer layer in the mBART$_{\text{large}}$ model. Language adapters are appended by each decoder layer. We follow the setting of language adapter used in (Pfeiffer et al., 2020b) while moving the layer normalization to the end of the adapter. For all experiments, we set d_a as 1024 and k as 6.

We utilize the Adam optimizer with learning rate scheduling. The warm-up step is 10000, and linear learning weight decay is used in the remaining training. We select the maximum learning rate from $\{1e-4, 3e-5\}$ according to the best result on the evaluation set. Decoding is done with beam search (beam size = 5) and length penalty ($\alpha = 1.5$ for text summarization and $\alpha = 3$ for question generation).

Adversarial-based method. The main idea is to use adversarial training to force the output of the

TM to be language-agnostic. Specifically, we introduce a language classifier to judge whether or not a text is from the source language. Given the TM's output h_L of a text x_L, the classifier calculates the probability that x_L belongs to the source language L_{src} as $\hat{y} = x_L \mathbf{W}_c^T$, where $\mathbf{W}_c \in \mathbb{R}^{d_a \times 1}$ is the weight of the classifier. We encourage the classifier to recognize x's language identity by minimizing a cross-entropy:

$$\mathcal{L}_{cls} = -\mathbb{I}_{x \in L_{src}} \cdot \log(\hat{y}) - (1 - \mathbb{I}_{x \in L_{src}}) \cdot \log(1 - \hat{y}),$$

where $\mathbb{I}_{x \in L_{src}} = 1$ when x is from the source language, otherwise 0. On the other hand, we encourage the TM to fool the language classifier:

$$\mathcal{L}_{adv} = -\mathbb{I}_{x \in L_{src}} \cdot \log(1 - \hat{y}) - (1 - \mathbb{I}_{x \in L_{src}}) \cdot \log(\hat{y}).$$

Besides, we utilize the cross-entropy loss between the decoder's output and the target sequence:

$$\mathcal{L}_{gen} = -(1 - \epsilon) \log p(i) - \sum_{j \neq i \in V} \frac{\epsilon}{|V| - 1} \log p(j)$$

The final loss is,

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{adv} + \mathcal{L}_{gen}$$

Note that the adversarial training needs data from the target language. As the annotated data from the target language can not be accessed, we leverage monolingual data.

Using multi-lingual representations. Like LAFT, we also need to provide language information to TM's output. Given the TM's output h_L^i and the encoder's output e_L^i, a gated mechanism aggregates h_L^i and e_L^i via a weighted sum as

$$\begin{aligned} \alpha^i &= \text{sigmoid}(\mathbf{W}_g([h_L^i, e_L^i]) + \mathbf{b}_g) \\ g_L^i &= \alpha^i h_L^i + (1 - \alpha^i) e_L^i \end{aligned}$$

where $\mathbf{W}_g \in \mathbb{R}^{d_h + d_a}$. Unlike the fusion mechanism, the gated mechanism is trained along with the whole model.