Towards Low-resource Language Generation with Limited Supervision

at The Big Picture Workshop, EMNLP 2023

Kaushal Kumar Maurya and Maunendra Sankar Desarkar

NLIP Lab, IIT Hyderabad, India

cs18resch11003@iith.ac.in

Introduction

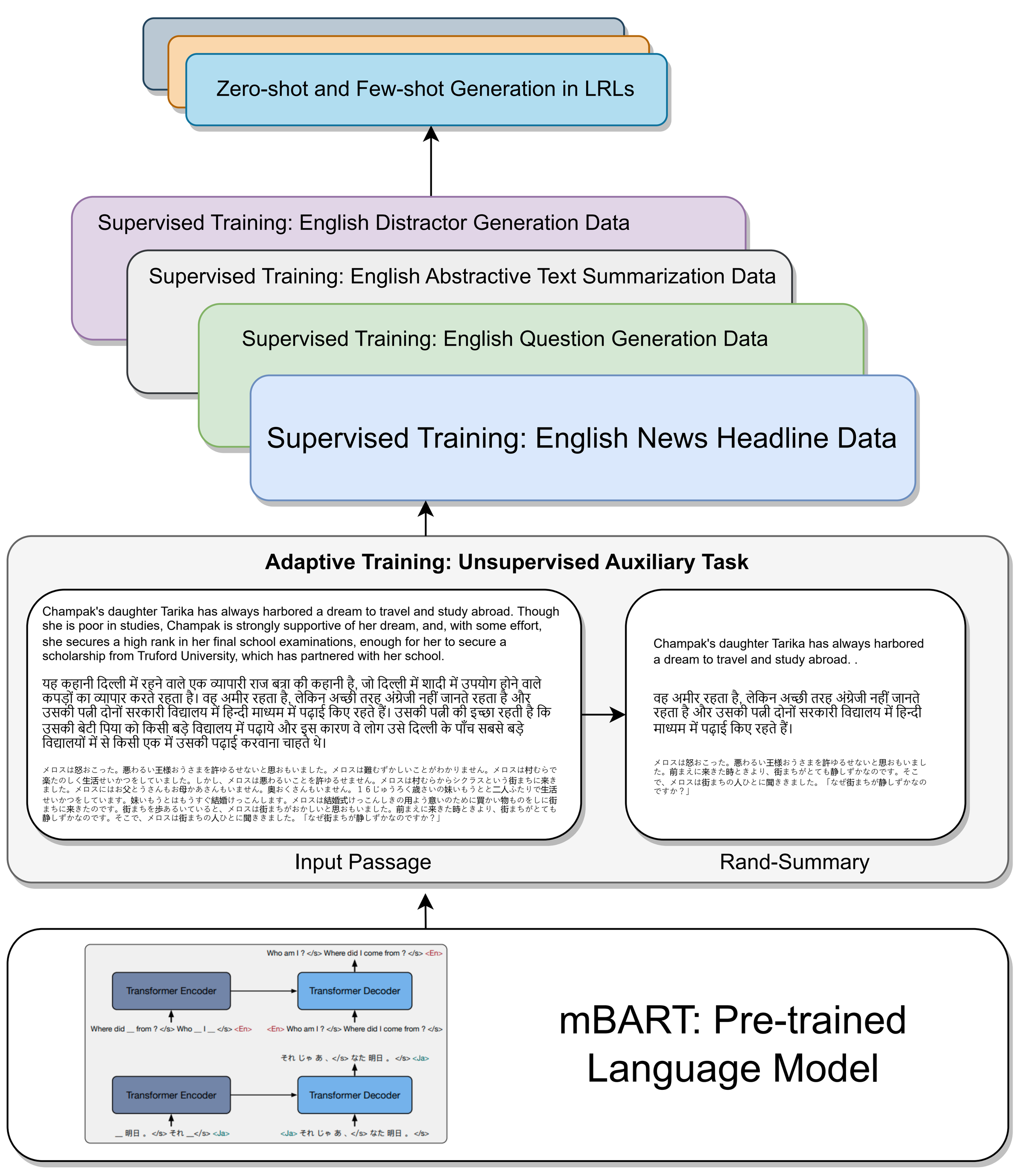
- There are 7000+ languages across the globe.
- The majority of NLP research focuses on English [1, 2] only - less inclusive and less diverse.
- The majority of the global population—roughly 95%—does not speak English as their primary language, and a staggering 75% do not speak English at all.
- Approximately 88% of languages are untouched by language technology [2].
- This thesis narrative is a step towards enabling language technology for low-resource languages (LRLs), specifically focused on NLG tasks.

Contributions

1. We proposed the **ZmBART** framework [3] to mitigate the catastrophic forgetting (CF) issues and enable well-formed zero-shot text generation in low-resource languages (LRLs).
2. We introduced the first meta-learning approach for cross-lingual transfer in LRLs (META-XNLG; [4]). It is based on language clustering to improve cross-lingual transfer, even for distant LRLs.
3. We presented a character span noise augmentation-based model (CHARSPAN; [5]) to enable machine translation for extremely low-resource languages (ELRLs).

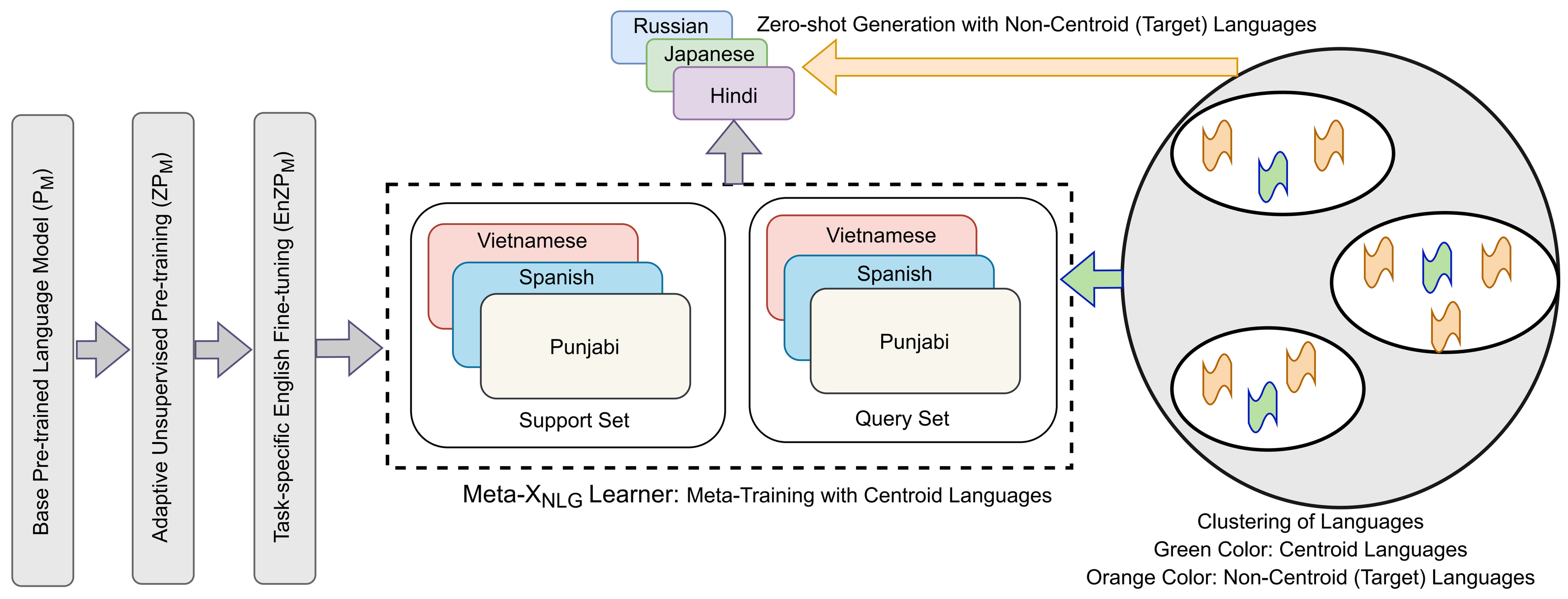
ZmBART: Mitigating Catastrophic Forgetting to Enable Zero-shot Language Generation

- **Zero-shot Cross-lingual Modeling:**
  - *Training with HRLs:* Train (fine-tune) a model (PLM) using a large annotated dataset from high-resource languages (HRLs), typically English. For instance, train with the English Abstractive Text Summarization (ATS) dataset.
  - *Zero-shot generation in LRLs:* Utilize the trained model for zero-shot inference. For instance, when given input in an LRL (e.g., Hindi), the model generates a summary in the same LRL (Hindi).
- **Catastrophic Forgetting Problem:**
  - After fine-tuning with task-specific HRL data, the model forgets the previous multilingual pre-training.
  - While attempting zero-shot generation in LRL, output comes in HRL, or code-mixed with HRL and LRL.
- **Proposed Approach:**
  - (1) Unsupervised adaptive training with an auxiliary task, i.e., RAND-SUMMARY objective.
  - (2) Adding a language tag, i.e., <fxx><2xx>. <xx>: ISO-2 language code.
  - (3) Freezing model components, i.e., freezing all the parameters of all word embedding and all decoder layers.
  - **RAND-SUMMARY:** It is a task of randomly predicting 10% of sentences from input passages. Requires only monolingual data in LRLs.
  - All three points above are necessary to mitigate CF and enable well-formed zero-shot generation in LRLs.
- We have evaluated the model across 3 LRLs and 4 NLG tasks on 4 datasets.



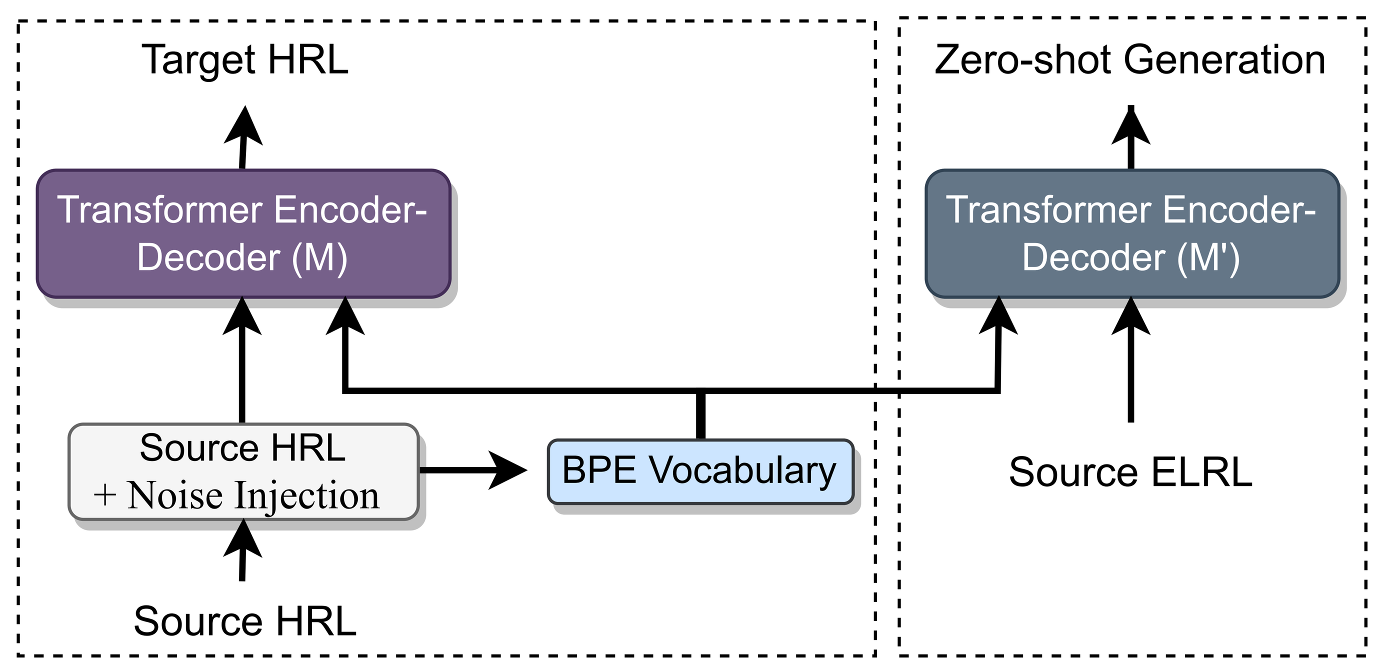
META-XNLG: Meta-Learning Approach to Improve Zero-shot Language Generation

- Cross-lingual modeling is a promising direction. However, the supervision transfer from HRL is uneven across LRLs, i.e., LRLs which are similar to HRL perform with high efficiency, and vice versa.
- Also, models do not account for cultural and linguistic aspects in the modeling.
- These factors lead to large *performance gaps* for LRLs.
- To the best of our knowledge, this is the first effort to use Meta-learning and Language clustering to uniformly transfer supervision for zero-shot generation.
- **Proposed Approach:**
  - Consider 30 languages and cluster them to find centroid and non-centroid languages.
  - Train a meta-learning algorithm with centroids and perform Zero-shot evaluation with non-centroid LRLs.
  - This enables *intra-cluster* and *inter-cluster* generalization to transfer supervision more uniformly.
- The evaluations are done across 30 LRLs, 5 datasets, and two NLG tasks.



CHARSPAN: Utilizing Lexical Similarity to Enable Zero-Shot MT for Extremely LRLs

- Large number of languages lack parallel data, have lack monolingual data, no representations in existing multilingual PLMs, called Extremely Low Resource Languages or ELRLs.
- **Task:** Machine Translation (MT) from ELRLs to English.
- **Proposed Approach:**
  - We propose a noise augmentation-based approach to enable cross-lingual transfer from HRL to *closely-related* LRLs.
  - We augment the character-span noise in the HRL side of the HRL-English parallel dataset to create a proxy training dataset.
  - Noise augmentation operations are: insert and delete; percentage: 9%-11%.
  - Training only with proxy HRL parallel data and evaluating with unseen ELRLs (zero-shot setting).
  - The noise augmentation acts as a regularizer and enables effective cross-lingual transfer to ELRLs.
- Evaluations are done with three typologically diverse language families across 12 ELRLs.



(a) Supervised Training Phase (b) Generation Phase

Conclusions

- We present three research efforts to enable language technology for LRLs (languages with limited data), with a special focus on NLG tasks.
- We hope that these collective efforts in a student thesis will advance the low-resource language generation space and be widely applicable for the general population.
- In the future, our aim is to develop a more unified modeling framework for the next 7000+ LRLs.

References

[1] Emily M Bender. The# benderrule: On naming the languages we study and why it matters. *The Gradient*, 14, 2019.

[2] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *ACL*, Online, 2020.

[3] Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. ZmBART: An unsupervised cross-lingual transfer framework for language generation. In *Findings of ACL*, pages 2804–2818, Online, August 2021.

[4] Kaushal Maurya and Maunendra Desarkar. Meta-x_{NLG}: A meta-learning approach based on language clustering for zero-shot cross-lingual transfer and generation. In *Findings of the ACL 2022*, pages 269–284, Dublin, Ireland, May 2022.

[5] Kaushal Kumar Maurya, Rahul Kejriwal, Maunendra Sankar Desarkar, and Anoop Kunchukuttan. Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages. *arXiv preprint arXiv:2305.05214*, 2023.

Acknowledgements

We thank all the human evaluators for evaluation, anonymous reviewers, and meta-reviewers for constructive feedback. We also thank the Big Picture Workshop for providing support to attend the conference.