Extending Generative NLP: Incorporating Diversity, Context, and Inclusivity in Neural Text Generation

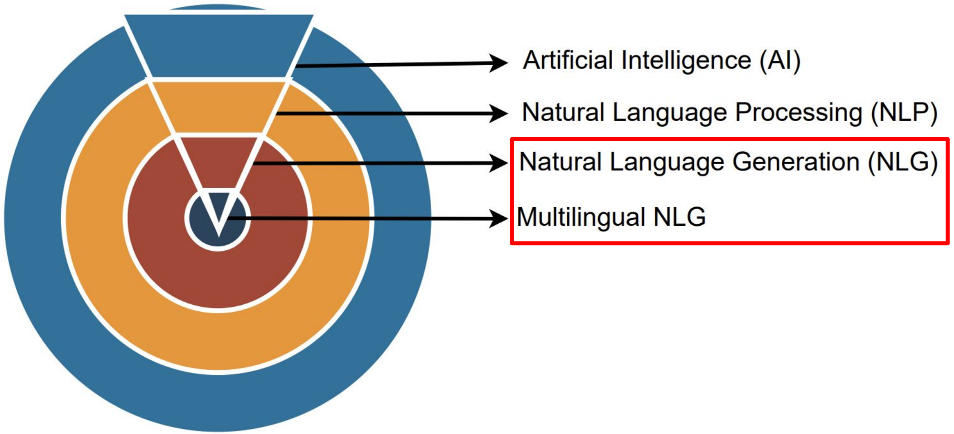
Kaushal Kumar Maurya

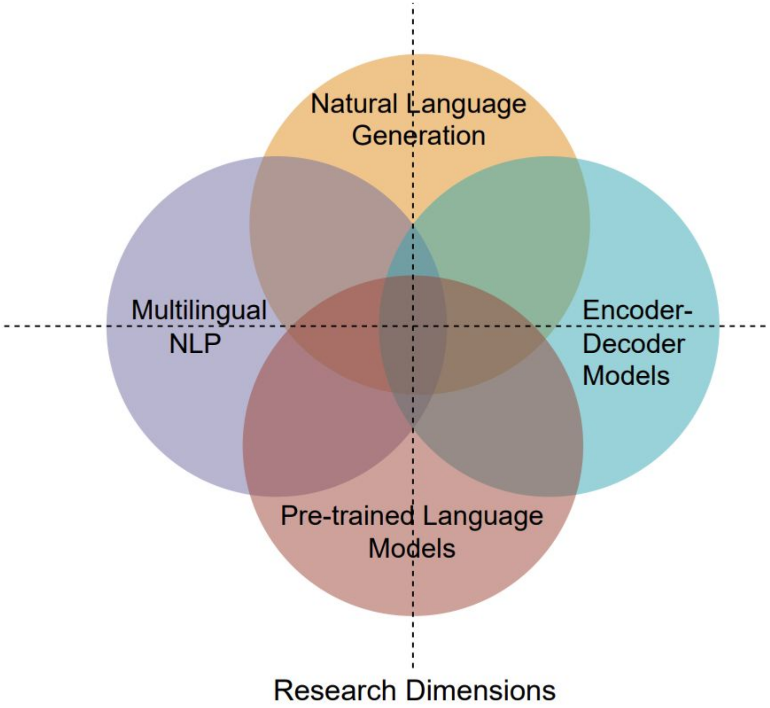
NLIP Lab, Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad, India

Outline

- Introduction
- Thesis Overview, Objectives and Contributions
- Deep Dive: Meta-Learning for Zero-Shot Cross-Lingual Generation
- Deep Dive: Machine Translation for Extremely Low-resource Languages
- Conclusion

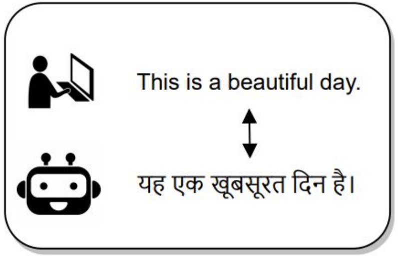
Introduction: Thesis Research Space



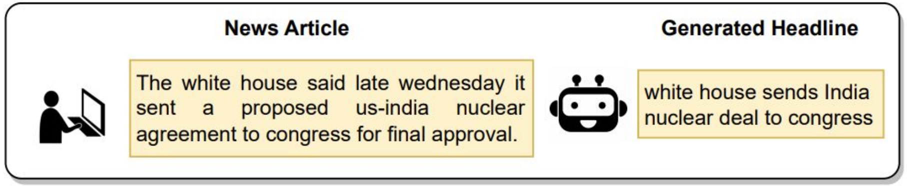


Introduction: Natural Language Generation (NLG)

NLG automates the generation of human-like text from a given input context.



Machine Translation



News Headline Generation

Image source: Internet

Need for further explorations: Thesis Objectives

Diverse Text Generation

Generates multiple outputs that are semantically related yet lexically diverse, all derived

- from a single input: one-to-many setup

Distractor generation for Reading Comprehension MCQs

Text Generation with Limited Context

Context that guides the generation is limited, making it challenging to generate relevant output

Autosuggest generation for incomplete search queries

Text Generation with Limited Labeled Data

Focusing on tasks/languages where available data is limited

- No task specific training data
- No parallel data and limited monolingual data

Multiple NLG tasks on multiple languages

Need for further explorations: Thesis Objectives

Diverse Text Generation

Generates multiple outputs that are semantically related yet lexically diverse, all derived

- from a single input: one-to-many setup

Distractor generation for Reading Comprehension MCQs

Definition: Diverse Text Generation

Diverse text generation is a *one-to-many* setup, where the model generates multiple outputs that are semantically related yet lexically diverse, all derived from a single input.

Passage	Ole bull was a very famous violinist from norway. He really liked to play the violin. But his father thought that playing the violin was not useful. So his father sent him to university to study. However, playing the violin was his dream. He did n't want to give up his dream. So he left university before he finished his studies and spent all his time and energy practicing the violin. Unfortunately, his violin teacher was not very good. So when it was time for him to start his concert tour, he still couldn't play the violin very well. Therefore, a milan newspaper critic criticized him and said that he was an untrained violinist. When facing this kind of problem, some people may become very angry and some people try to learn from it. Fortunately, ole bull belonged to the second group. He went to the newspaper office and found the critic. Instead of being angry, he talked about his mistakes with the man and listened to the man's advice. After he met the critic, he gave up the rest of his concerts. Then he went back to practice the violin with the help of good teachers. In the end, he got great success when he was only 26. He also became one of the most famous violinists in the world.
Question	Why didn't ole bull's father like him to play the piano?
Correct Answer	Because he thought playing the violin was useless.
Distractor - I	Because playing the violin would cost lots of money.
Distractor - II	Because the violin was not good.
Distractor - III	Because he didn't like to play the violin.

Distractor Generation

News	China filed the highest number of patent applications globally in 2020 retaining its top position for the second consecutive year the UN s World Intellectual Property Organization WIPO said. China filed 68,720 applications last year while the US filed 59,230. In 2019 China had replaced the US as the top patent application filer for the first time in over four decades.		
Reference Headlines	China files highest patents globally for 2nd year in a row: UN	China becomes world's top patent filer after four decades with US on top	China extends lead over U.S. in global patents filings U.N. says

Diverse News Headline Generation

Definition: Distractor Generation

Given a reading comprehension MCQ, i.e., <passage, question, correct answer> triplet, it is the task of generating multiple incorrect options, i.e., distractors.

Ideal distractors should be:

1. Contextually related with question
2. Semantically dissimilar with answer
3. Diverse from each other
4. Confusion-inducing

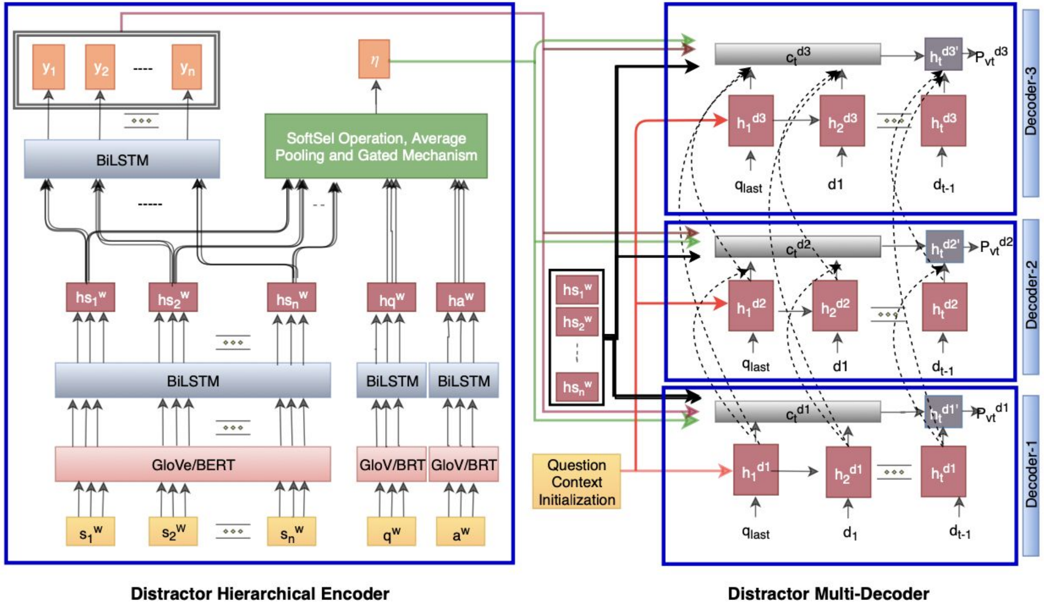
Distractor Generation: Contribution

Encoder Side:

- SoftSel operation
- Gated Mechanism
- Semantic Decoupling

Decoder Side:

- Multi-decoder Network
- Question Context Initialization
- Decoder interactions



HMD-Network

[1]. Kaushal Kumar Maurya and Maunendra Sankar Desarkar. "Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors for multiple-choice questions for reading comprehension." *CIKM* 2020.

Need for further explorations: Thesis Objectives

Diverse Text Generation

Generates multiple outputs that are semantically related yet lexically diverse, all derived

- from a single input: one-to-many setup

Distractor generation for Reading Comprehension MCQs

Text Generation with Limited Context

Context that guides the generation is limited, making it challenging to generate relevant output

Autosuggest generation for incomplete search queries

Personalized Query Auto-Completions (PQAC)

Generating top-m completions given the user-specific <session, prefix>

- Query Auto Completion (QAC): Recommend a list of relevant complete queries for partially typed search query (i.e. prefix)
- Helps in:
  - Saving keystrokes
  - System's understanding of search intent
  - Assisting users in efficiently expressing their intent

Session: mountains images||caves images||mountainside caves||mountain caves||timber wolves
Prefix: wolf p
Correct Query: wolf poetry

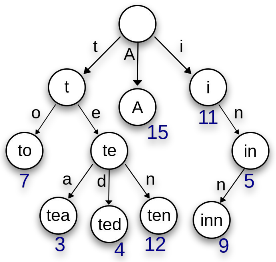
Generations:
1. wolf poetry
2. wolf pictures
3. wolf photos
4. wolf pics
5. wolf picture
6. wolf photo
7. wolf prints
8. wolf print

Sample from Bing Query log

Credit: Bing query log

Limited Context in Existing PQAC Models

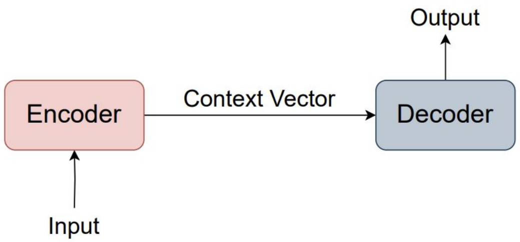
Trie/Ranking Models



\+ Suggestions are more meaningful as they come from user log

\- No personalization
\- Provide limited number of suggestions
\- No suggestions for unseen prefixes

NLG Models

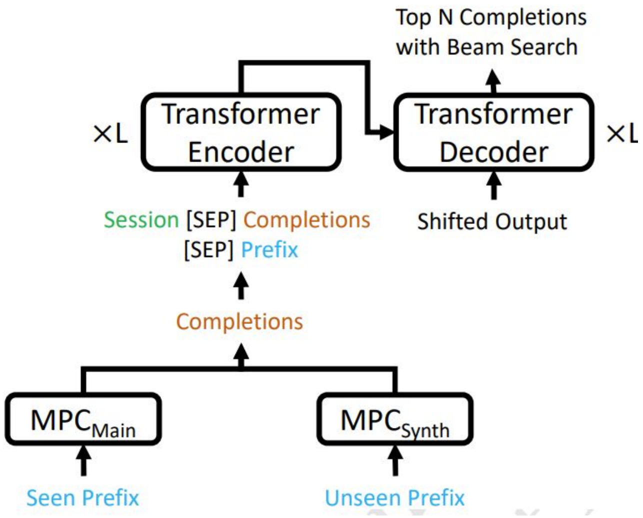


\+ Can model personalization
\+ Generate suggestions for unseen prefixes

\- For short-prefixes, suggestions are bad due to limited context
\- Learn the generation biases from training dataset

Limited Context in Existing PQAC: Contribution

- Best of both (Trie and NLG): Augment top trie completions in NLG model for short and unseen prefixes
- Proposed Retrieval-Augmented Generation (RAG) type modeling framework
- This mitigates the issues of Limited Context in PQAC and boost the model performance
- This is first such study



Trie-NLG Model

[2]. Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Manish Gupta, and Puneet Agrawal. "trie-nlg: trie context augmentation to improve personalized query auto-completion for short and unseen prefixes." *ECML-PKDD (DAMI) 2023*

Need for further explorations: Thesis Objectives

Diverse Text Generation

Generates multiple outputs that are semantically related yet lexically diverse, all derived

- from a single input: one-to-many setup

Distractor generation for Reading Comprehension MCQs

Text Generation with Limited Context

Context that guides the generation is limited, making it challenging to generate relevant output

Autosuggest generation for incomplete search queries

Text Generation with Limited Labeled Data

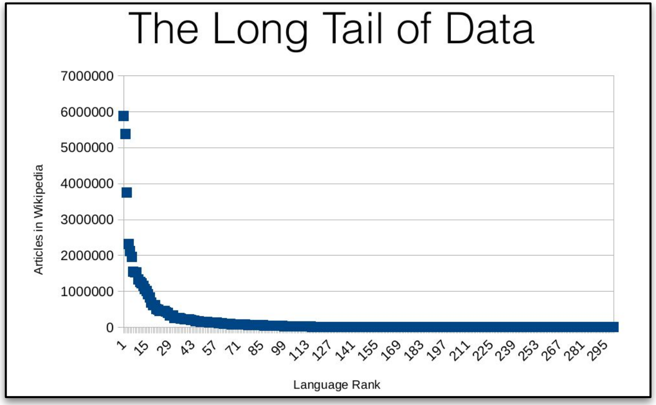
Focusing on tasks/languages where available data is limited

- No task specific training data
- No parallel data and limited monolingual data

Multiple NLG tasks on multiple languages

Landscape of Low-resource Languages (LRLs)

- 7000+ languages across the globe [3]
- Only ~300 languages has wikipedia page
- The majority of NLP research focuses on English [3, 4] 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[1]



Source: Graham Neubig Multilingual NLP Lectures

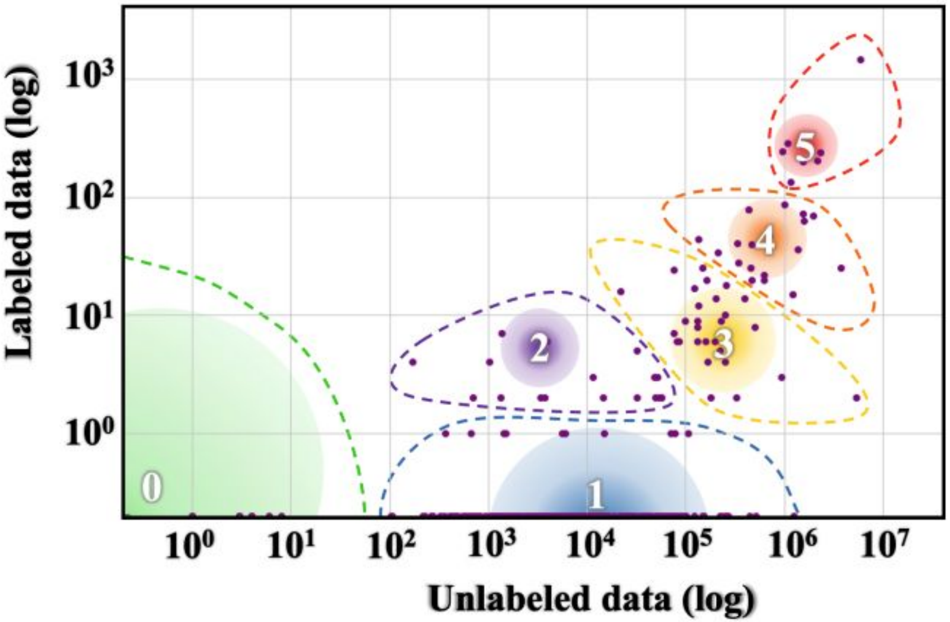
[3] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 2020 6282–6293.

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

[1]https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

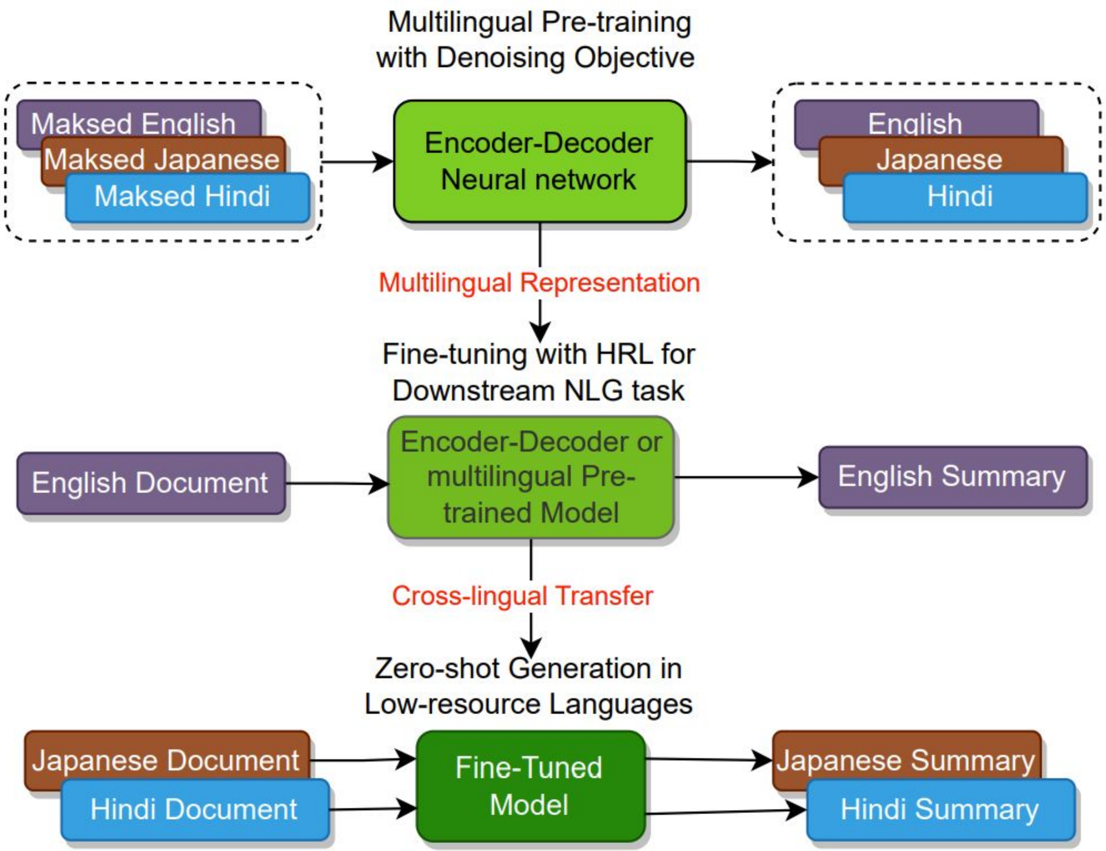
Limited data for LRLs

- 88% languages fall into class 0 and untouched by language technology [3]
- Task-specific NLG labeled data in LRL is even more rare
- Only ~100 languages are part of existing large language model, even for those languages, NLG adaptability is challenging [5]



[3] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In ACL 2020.
[5] Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. In EMNLP 2023.

Hopeful Direction: Cross-lingual Modeling in NLP



Cross-lingual Modeling: Challenges and Contributions

Challenge 1: Catastrophic Forgetting (CF) Problem

- After fine-tuning with task-specific HRL data, the model forgets the previous multilingual pre-training.
- While attempting zero-shot generation in LRL, the generation is in HRL, or code-mixed with HRL and LRL.

Our Proposal: ZmBART Framework

1. Proposed unsupervised adaptive pre-training-based framework -ZmBART
2. Only require small monolingual data in LRLs
3. Successfully mitigate the the is CF problem and generates well-formed zero-shot generation in LRLs
4. Evaluated across 4 tasks and 3 languages
5. In zero-shot and few-shot settings

[6]. Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha."ZmBART: An Unsupervised Cross-lingual Transfer Framework for Language Generation." In Findings of the Association for Computational Linguistics: ACL 2021, pages 2804–2818.

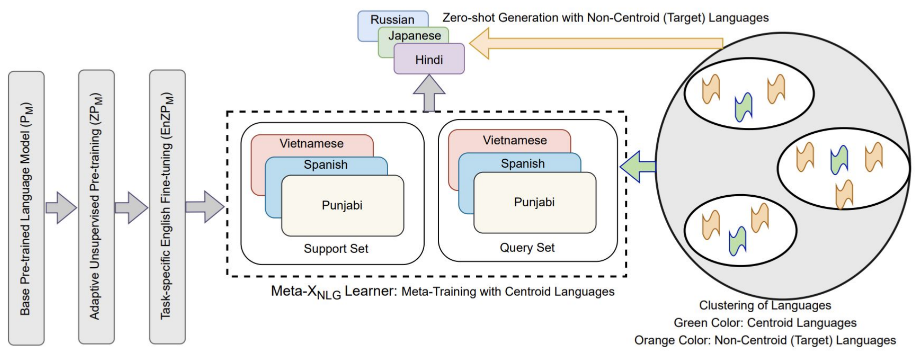
Cross-lingual Modeling: Challenges and Contributions

Challenge 2: Non-Uniform Supervision Transfer

- Supervision transfer from HRL is uneven across LRLs, i.e., LRLs which are similar to HRL, the transfer is high, and vice versa.
- Models do not account for cultural and linguistic aspects in the modeling.
- These factors lead to large performance gaps for LRLs.

Our Proposal: Meta-XNLG Framework:

- First study to propose modeling framework to transfer cross-lingual signals more uniformly with Meta-learning (MAML) and Language clustering



[7]. Kaushal Kumar Maurya and Maunendra Desarkar. “Meta-XNLG: A Meta-Learning Approach Based on Language Clustering for Zero-Shot Cross-Lingual Transfer and Generation.” In Findings of the Association for Computational Linguistics: ACL 2022, pages 269–284, Dublin, Ireland.

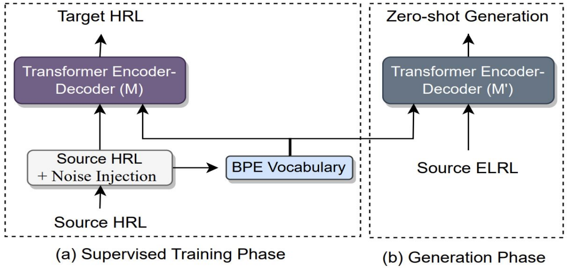
Cross-lingual Modeling: Challenges and Contributions

Challenge 3: Machine Translation for Extremely LRLs (ELRLs)

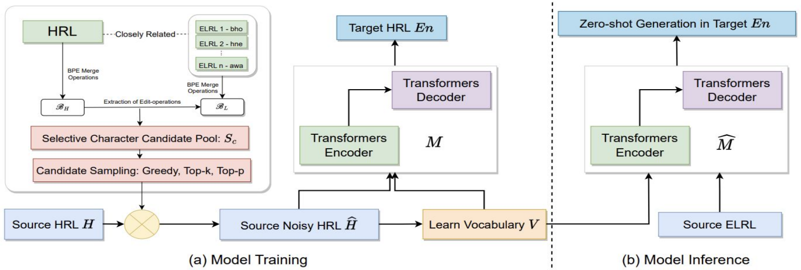
Lack parallel data, have limited monolingual data, and language representations absent in mPLMs.

Our Proposal: SelectNoise [8] and CharSpan [9] Models

- Based on noise augmentation
- Small monolingual data (1k examples): SelectNoise; No monolingual data: CharSpan



CharSpan Model

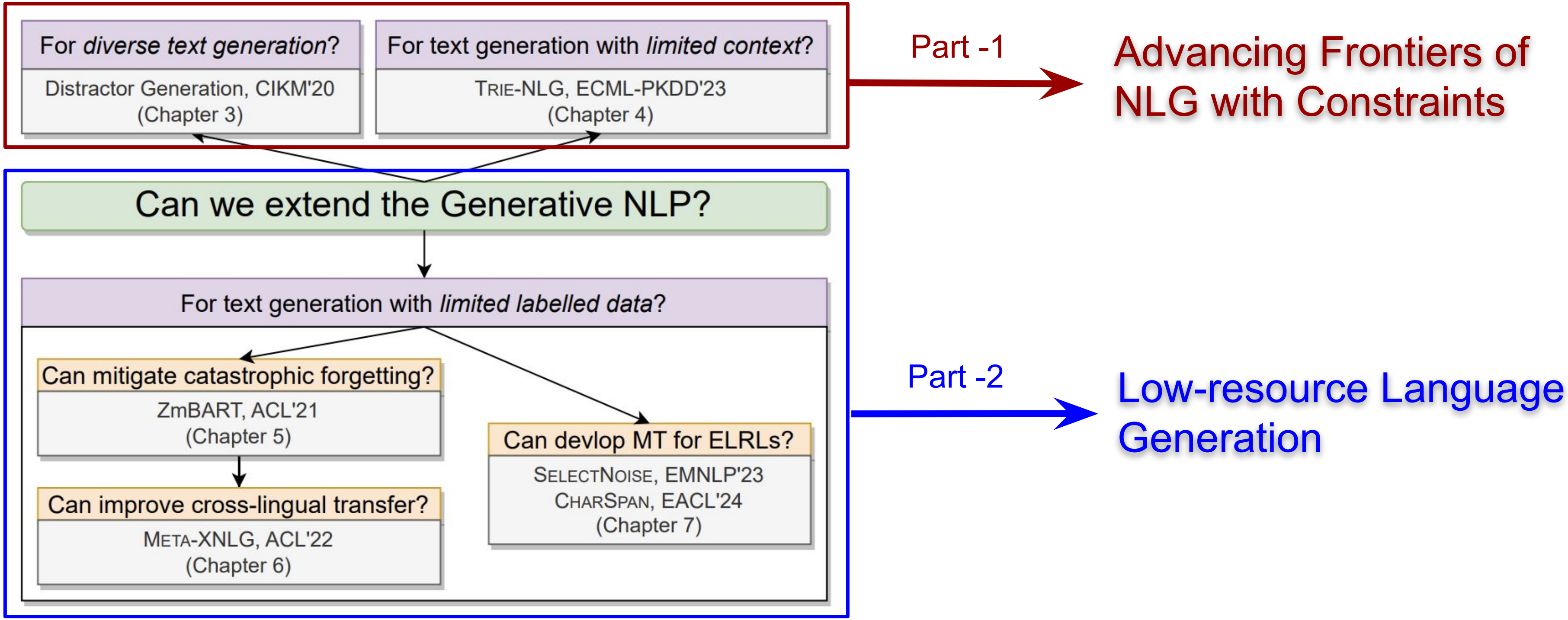


SelectNoise Model

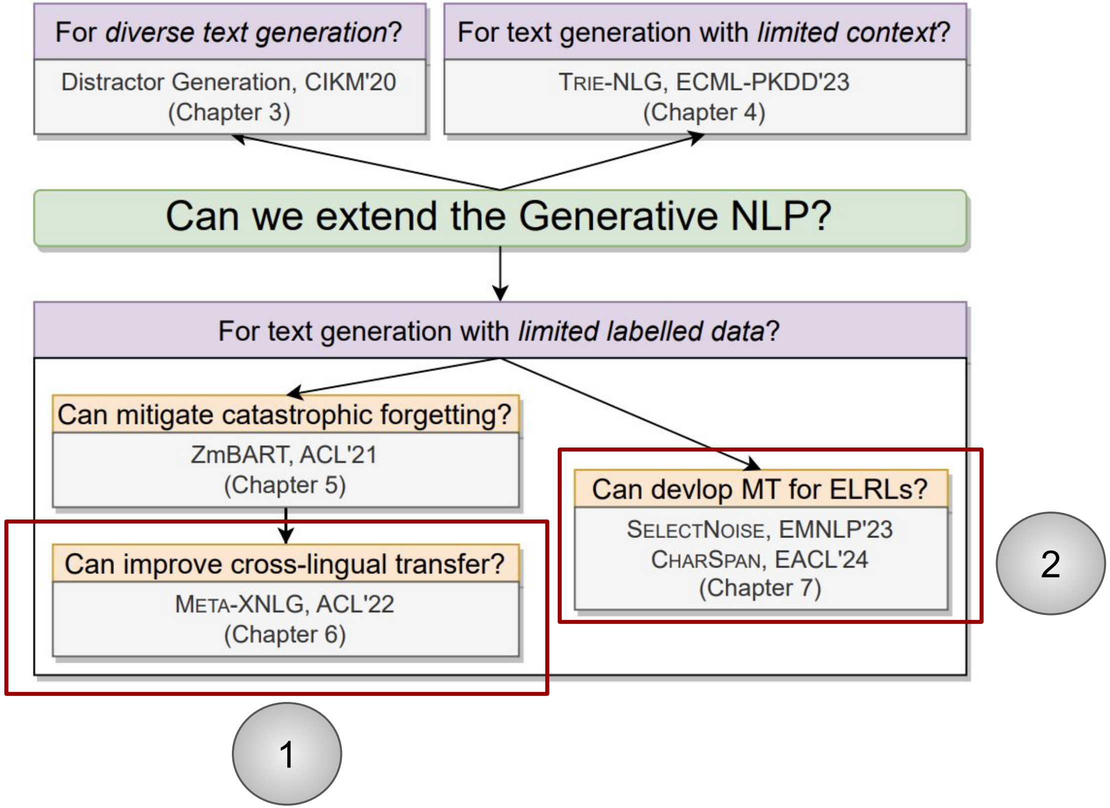
[8] Maharaj Brahma*, Kaushal Kumar Maurya*, and Maunendra Desarkar. "SelectNoise: Unsupervised Noise Injection to Enable Zero-Shot Machine Translation for Extremely Low-Resource Languages." In Findings of EMNLP 2023.

[9] Kaushal Kumar Maurya, Rahul Kejriwal, Maunendra Desarkar and Anoop Kunchukuttan. "CharSpan: Utilizing Lexical Similarity to Enable Zero-Shot Machine Translation for Extremely Low-Resource Languages." In EACL 2024.

Thesis Organization



Deep Dive: Meta-XNLG and SelectNoise

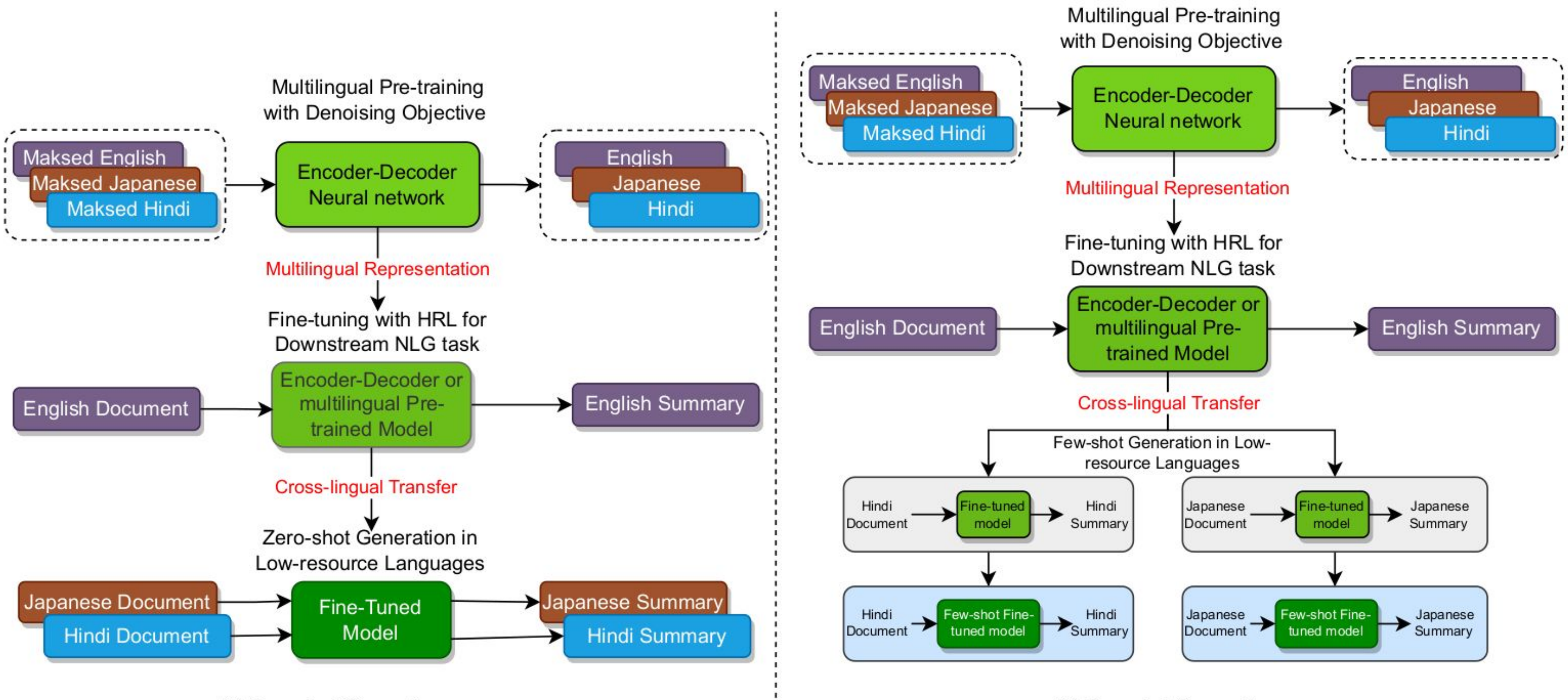


Introduction: Terminology

- Zero-shot Generation: An NLG model to generate output text in a language L (or domain D) without prior explicit labeled training in L (or D).
- Few-shot Generation: An NLG model to generate output text in a language L (or domain D) with limited labeled training examples in L (or D).
- Low-Resource Language Generation: It is the task of generating textual output from a given task with a limited amount of training data or linguistic resources.
- Cross-Lingual Transfer and Generation: It is a task in which a model learns a generative task from labeled data in one language (typically English) and then performs the equivalent generative task in another language [10].

[10] Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. Overcoming catastrophic forgetting in zero-shot cross- lingual generation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9279–9300, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Introduction: Terminology [Cont...]



(a) Zero-shot Generation

(b) Few-shot Generation

NATURAL LANGUAGE
& INFORMATION
PROCESSING LAB

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Meta-XNLG: A Meta-Learning Approach Based on Language Clustering for Zero-Shot Cross-Lingual Transfer and Generation

Kaushal Kumar Maurya and Maunendra Sankar Desarkar

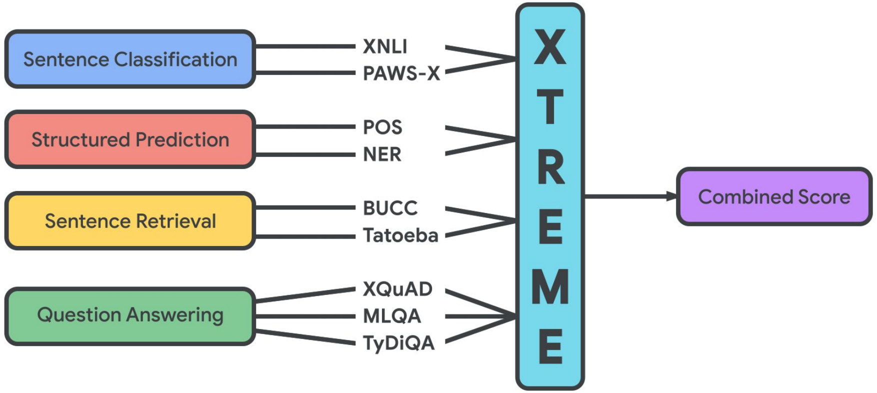
Natural Language & Information Processing Lab (NLIP Lab)
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad, India

❑ Meta-XNLG: Introduction

Efforts in Cross-lingual Research

Rank	Model	Submission Date	Dataset Translation	Parameter (Million)	QG	NTG	XGLUE-Generation Score
1	**SAG** (MS Ads Creative)	2022-01-12	Yes	270M	12.6	12.4	12.5
2	ProphetNet-X (USTC+MSRA)	2021-04-22	No	270M	12.2	11.6	11.9
3	**Unicoder Baseline** (XGLUE Team)	2020-05-25	No	270M	10.6	10.7	10.7
4	**MP-Tune** (ByteDance MLNLC)	2021-01-12	Yes	304M	8.1	9.4	8.7

XGLUE-Benchmark [11]



XTREME-Benchmark [12]

[11] Liang, Yaobo et al. “XGLUE: A New Benchmark Datasetfor Cross-lingual Pre-training, Understanding and Generation.” *Conference on Empirical Methods in Natural Language Processing* (2020).
[12] Hu, Junjie, et al. "Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation." International Conference on Machine Learning. PMLR, 2020.

Current Limitations

- Non-uniform supervision transfer from HRL to LRLs
  - More similar LRL to the HRL → better supervision transfer and vice-versa.
  - Cultural and linguistic aspects are not considered in the modeling [13, 14]
  - Leads to large performance gaps.

[13] Guokun Lai, Barlas Oguz, Yiming Yang, and Veselin Stoyanov. 2019. Bridging the domain gap in cross-lingual document classification. CoRR, abs/1909.07009.

[14] Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Hopeful Direction

- Learn shareable structures across multiple tasks with limited annotated data: Modeling with Meta-Learning [15]
    - Constraint: all tasks should share some common structure (or come from a task distribution)
    - The world's different languages follow this constraint
    - As they came into existence with a common goal of communication, and share some structure.
- We consider all the languages as tasks

[15] Bengio, Yoshua, Samy Bengio, and Jocelyn Cloutier. Learning a synaptic learning rule. Université de Montréal, Département d'informatique et de recherche opérationnelle, 1990.

Hypothesis and Modeling Direction

Hypothesis: Meta-learning algorithm trained with few *typologically diverse* languages (as training task) provide language-agnostic initialization for the zero-shot cross-lingual generation.

- We propose Meta-XNLG, a framework for effective cross-lingual transfer and generation based on Model-Agnostic Meta-Learning (MAML; [16]) algorithm.
- This was first attempt to study meta-learning techniques for cross-lingual natural language generation (XNLG).
- Particularly, we focus on zero-shot XNLG for low-resource languages.

[16] Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." International conference on machine learning. PMLR, 2017.

❑ Meta-XNLG: Background

Meta-Learning

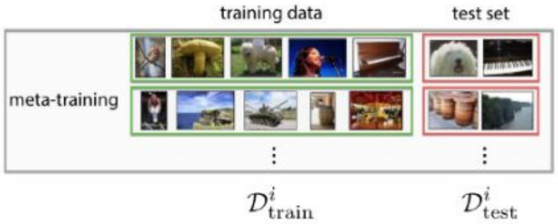
- Meta-learning or learning to learn [15]
  - A learning paradigm
  - Model is trained on diverse tasks (and learns structures)
  - Quickly adapts to new tasks given a handful of examples.

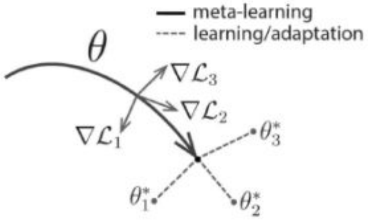
- Among others, we focus on optimization-based algorithms, i.e., Model Agnostic Meta-Learning (MAML) due to its recent success.

[15] Bengio, Yoshua, Samy Bengio, and Jocelyn Cloutier. Learning a synaptic learning rule. Université de Montréal, Département d'informatique et de recherche opérationnelle, 1990.

Model-Agnostic Meta-Learning (MAML)

a general recipe:





$$\theta \leftarrow \theta - \beta \sum_i \nabla_\theta \underbrace{\mathcal{L}(\theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}^i_{\text{train}}), \mathcal{D}^i_{\text{test}})}_{\text{"meta-loss" for task } i}$$

* in general, can take more than one gradient step here

** we often use 4 – 10 steps

Chelsea Finn

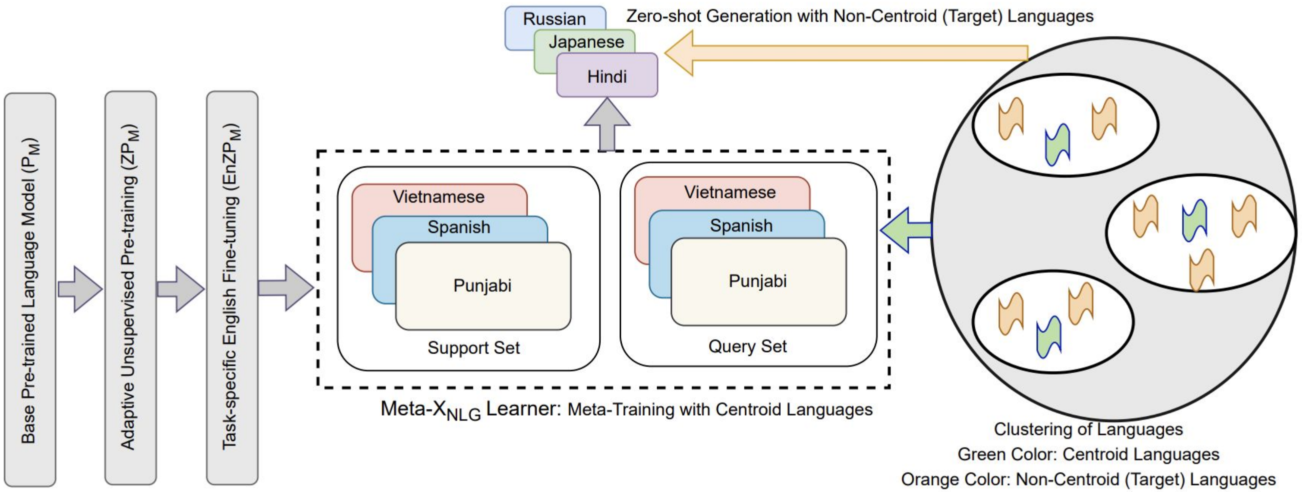
Finn et al., "Model-Agnostic Meta-Learning"

Slide Credit: Sergey Levine

[16] Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." International conference on machine learning. PMLR, 2017.

❑ Meta-XNLG: Proposed Approach

Meta-XNLG: Overview



1. Cluster the languages and find the centroid languages
2. Train MAML with centariod languages
3. Evaluate the Trained MAML model with non-centroid languages in zero-shot setting while overcoming accidental translation problem

Meta-XNLG: Intuition

- **Generalization Goals:** Intra-cluster and Inter-cluster
  - **Intra-cluster:** Training with one centroid language per cluster
  - **Inter-cluster:** Training with multiple centroid languages across cluster

- **Number of clusters (meta-training language) Vs Generalization**
  - → single cluster → over-generalization → fails to learn different typological structures
  - → too many clusters → many centroid languages → many typological structures → learning distracted
  - Empirical evidence: three clusters provide best generalization

Language Clustering: Representation of Languages

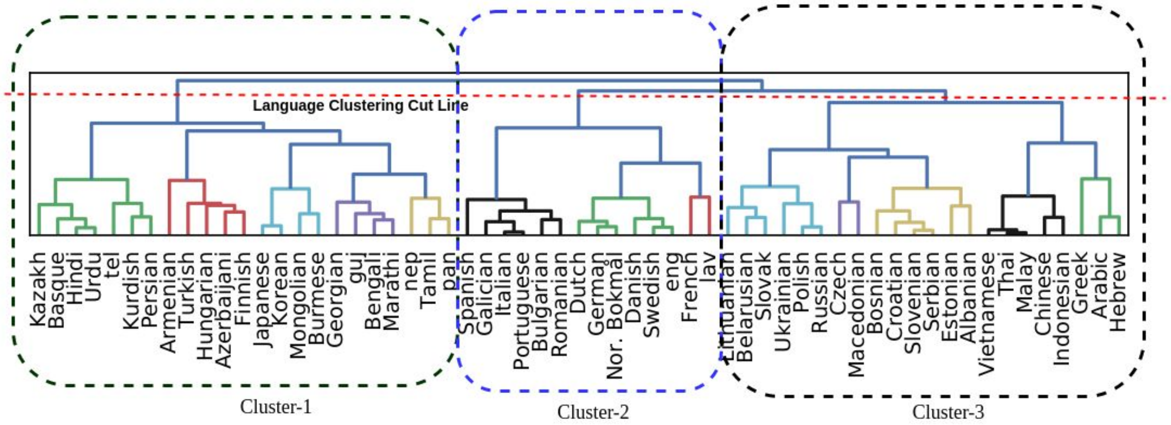
- **Typologically Learned:** Use typological information from linguistic knowledge-base like WALS [17]
- **Task-learned:** Extract learned language tag representations from tasks like machine translation [18]
- **Multi-View:** Fuse typologically learned and task-learned language representations using singular vector canonical correlation [19]

[17] Dryer, Matthew S., and Martin Haspelmath. "The world atlas of language structures online." (2013).

[18] Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In EMNLP 2017.

[19] Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In EMNLP 2020.

Language Clustering

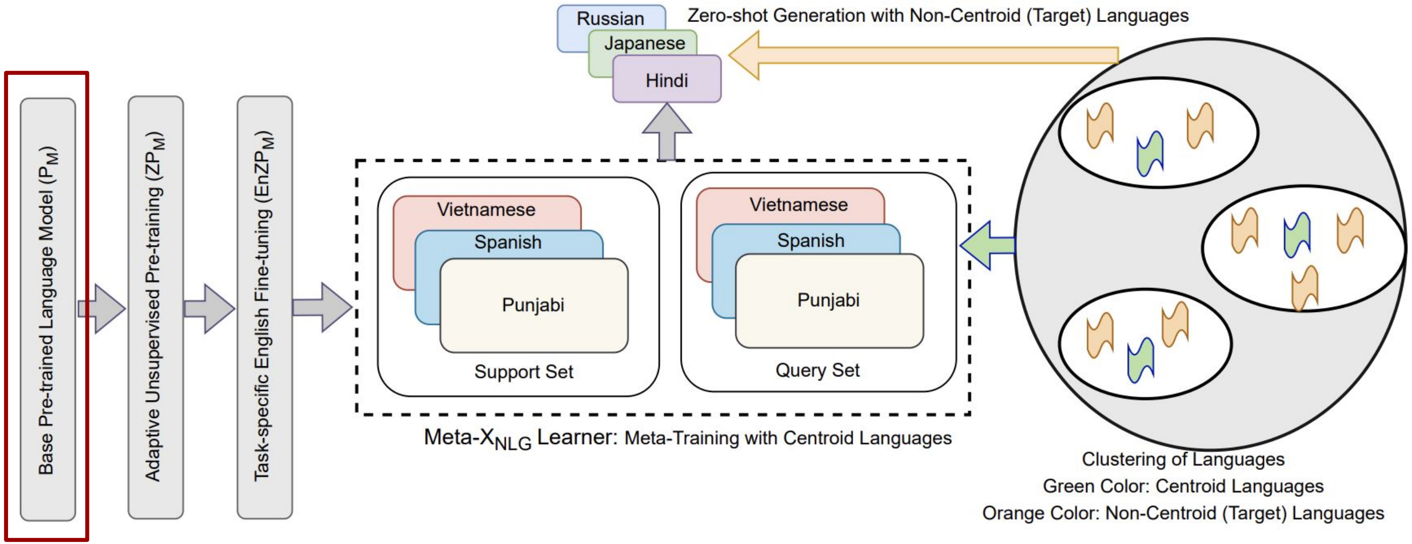


➢ Given cluster $C = \{L_1, L_2, \ldots L_t\}$ and d is cosine distance then the centroid language $L^* \in C$ is defined as:

$$L^* = \arg\min_{L_i \in C} \sum_{L_j \in C} d(L_j, L_i)$$

Cluster-1(14)	Cluster-2(8)	Cluster-3(8)
hi,ur,te,tr,ja,fi,ko,gu, bn,mr,np,ta,pa,sw	es,it,pt,ro, nl,de,en,fr	ru,cs,vi,th, zh,id,el,ar

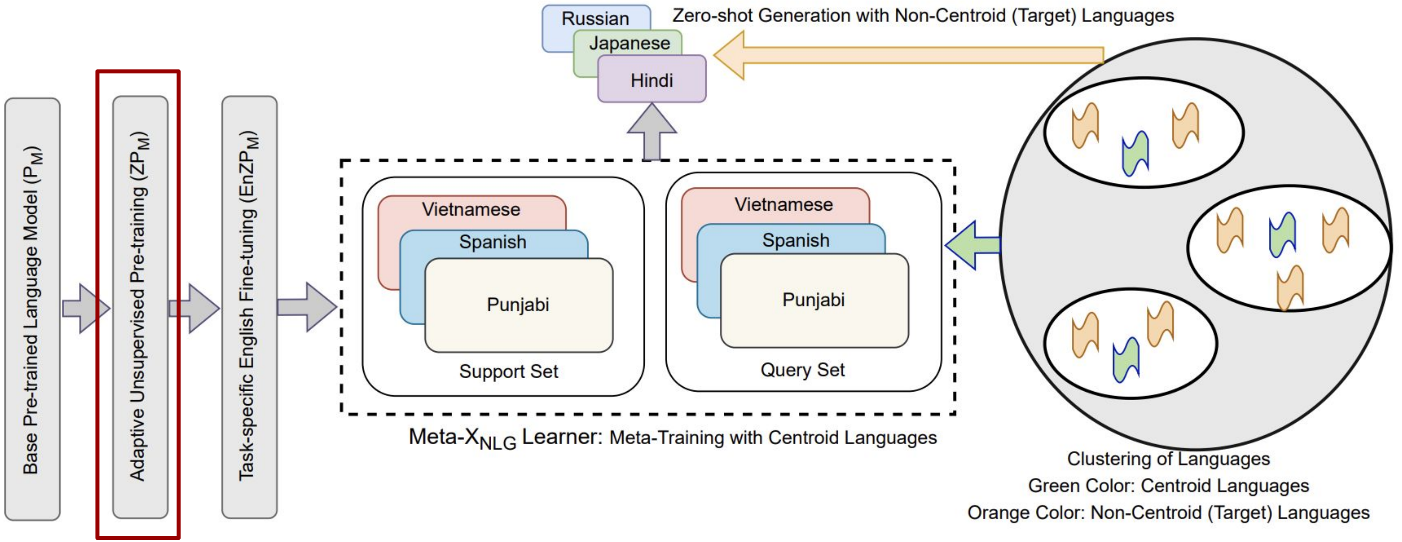
Model Training: Step-1



[Step-1] Selection of Base Pre-trained Model:

1. Pre-trained model-agnostic
2. Used sequence-to-sequence multilingual pre-trained language
3. We use mT5

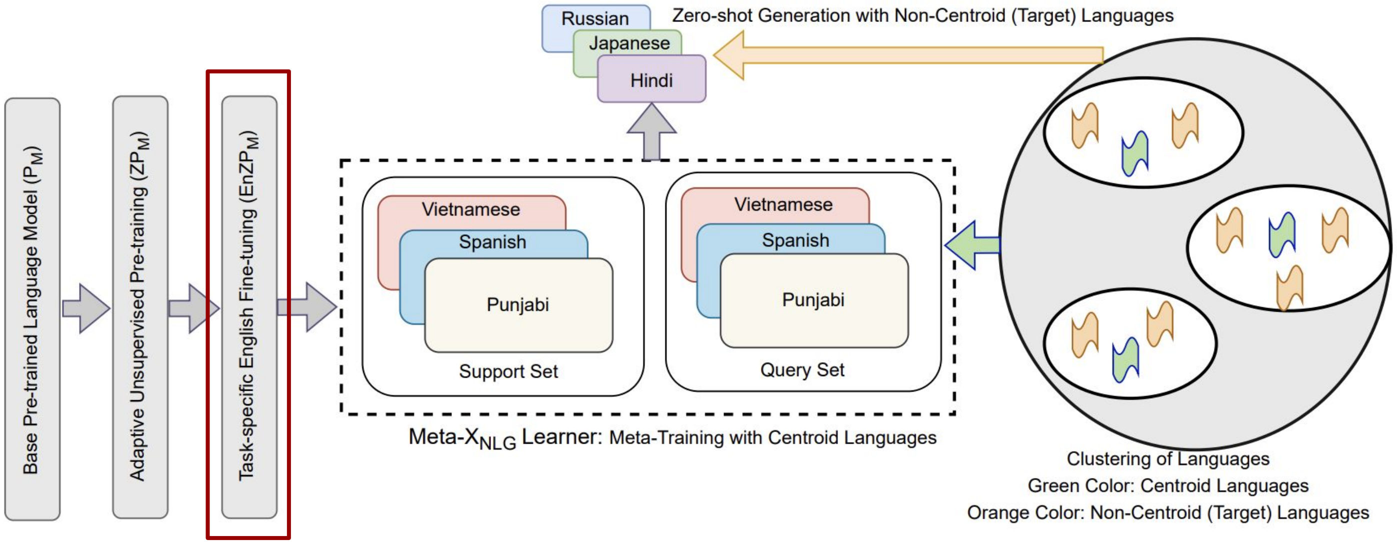
Model Training: Step-2



[Step-2] Adaptive Unsupervised Pre-training:

1. An adaptive pre-training step on top of mT5
2. Use rand-summary training objective (more on this later)
3. Helps in mitigate catastrophic forgetting/accidental translation/off-target problem

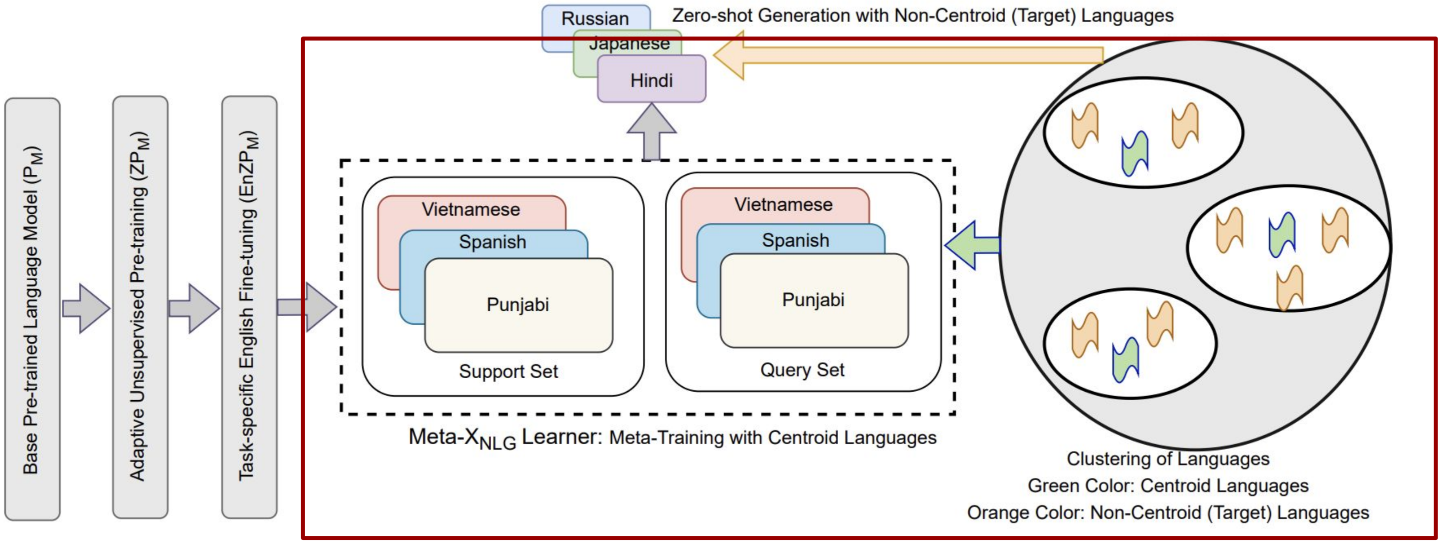
Model Training: Step-3



[Step-3] Fine-tuning with High Resource Language:

1. Task specific supervised training with HRL (English) dataset
2. The supervision will transfer from English to LRLs

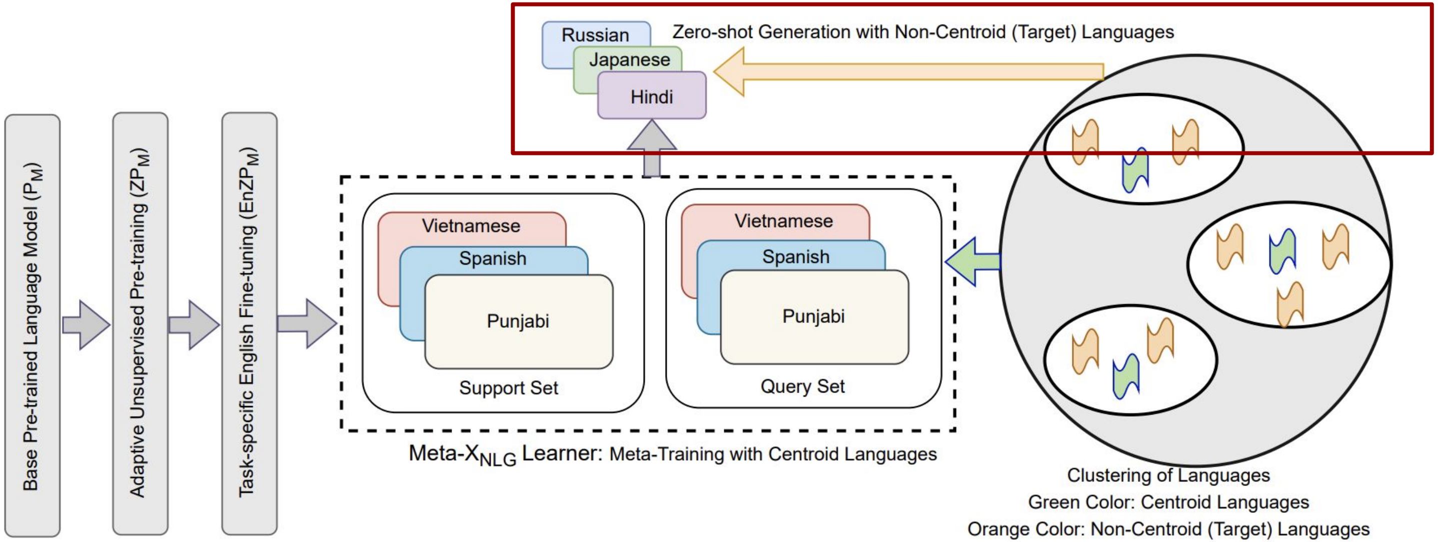
Model Training: Step-4



[Step-4] Meta-Training with Low-resource Centroid Languages:

1. Train MAML with centroid languages
2. Obtain meta-learned checkpoints

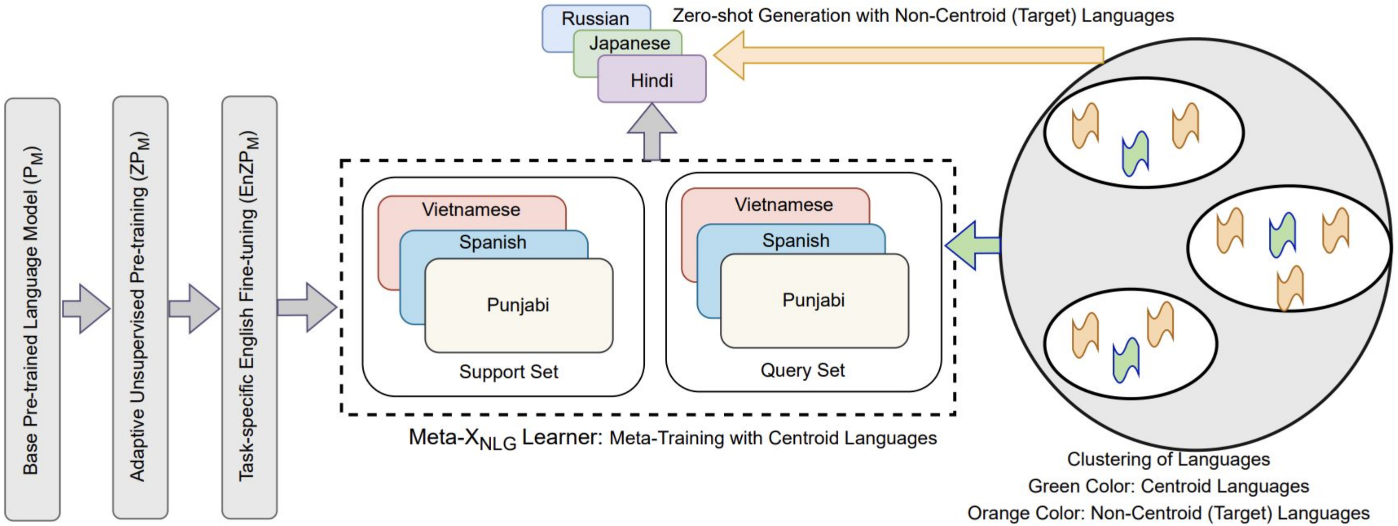
Model Training: Step-5



[Step-5] Meta-adaptation for Zero-shot Evaluation:

1. Evaluate the meta-learned model with non-centroid languages
2. Evaluations are done in the zero-shot setting

Meta-XNLG: Summary



1. [Step-1] Selection of Base Pre-trained Model
2. [Step-2] Adaptive Unsupervised Pre-training
3. [Step-3] Fine-tuning with High Resource Language
4. [Step-4] Meta-Training with Low-resource Centroid Languages
5. [Step-5] Meta-adaptation for Zero-shot Evaluation

Meta-XNLG: Algorithm

Algorithm 1 Meta Learning Algorithm

Require: Task set distribution $p(D)$, pre-trained model $EnZP_M$ (P) with parameters θ_P, meta-learner f_θ with parameter θ.

Require: α, β: step size hyper-parameters

Initialize $\theta \leftarrow \theta_P$

while not done **do**

  Sample batch of tasks $T = T_1, T_2, \ldots T_b \sim p(D)$

  for all T_i in T **do**

    Initialize $\theta_i \leftarrow \theta$

    Split D_i to form support set S_i and query set Q_i

    for all inner_iter steps m **do**

      Compute $\nabla_{\theta_i^{(m)}} L_{T_i}^{S_i}(P_{\theta_i^{(m)}})$

      Do SGD: $\theta_i^{m+1} = \theta_i^m - \alpha \nabla_{\theta_i^{(m)}} L_{T_i}^{S_i}(P_{\theta_i^{(m)}})$

    end for

    MetaUpdate: $\theta = \theta - \beta \nabla_\theta \sum_{j=1}^{b} L_{T_i}^{Q_i}(P_{\theta_i^{(m)}})$

  end for

end while

Do zero-shot/few-shot learning with meta-learner f_{θ^*} where θ^* is learned optimal parameters of meta-learner.

Catastrophic Forgetting / Accidental Translation/Off-trager Problem

- In zero-shot setting, model suffers from ill-formed generation for unseen low-resource and problem is known as catastrophic forgetting or accidental translation or off-trager problem

- Where model generates whole/part of the output in the fine-tuning HRL [20] i.e., HRL or code-mixed with HRL and LRL

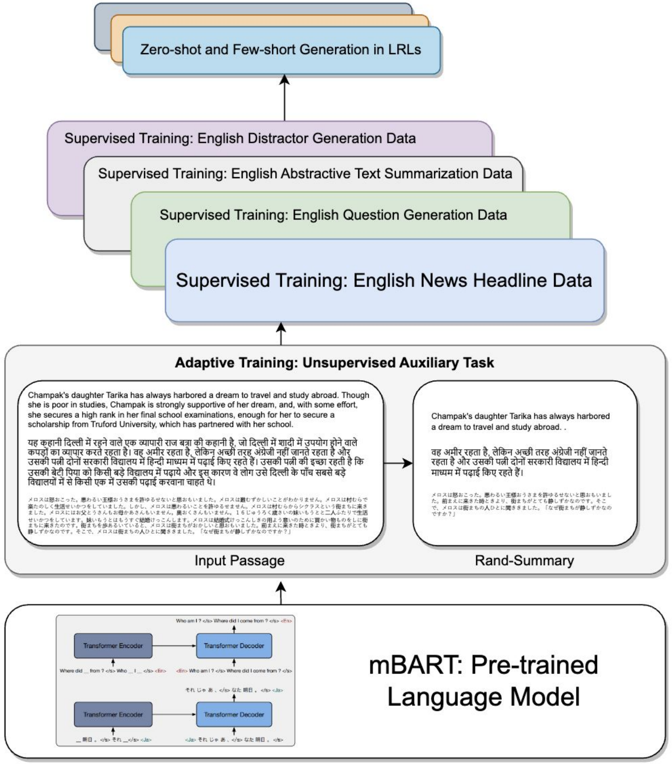
[20] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

❑ ZmBART Framework

[6] Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha."ZmBART: An Unsupervised Cross-lingual Transfer Framework for Language Generation." In Findings of the Association for Computational Linguistics: ACL 2021, pages 2804–2818, Online.

Overview of ZmBART Model

- ➢ It is an unsupervised *framework* to mitigate CF/AT/OT problem and cross-lingual transfer and generation
- ➢ Enable zero-shot well formed generation in target LRLs
- ➢ Developed on mBART pre-trained model checkpoint
- ➢ Evaluated with 4 NLG tasks with 3 languages (HRL: English and LRLs: Hindi, Japanese)



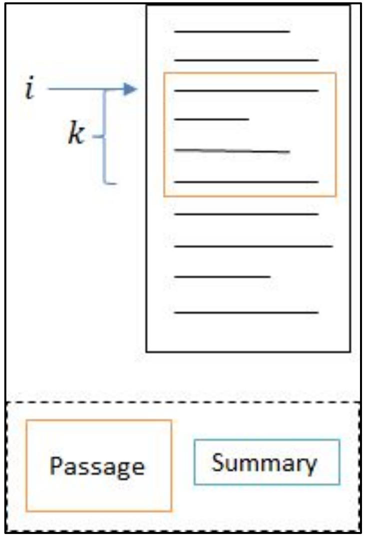
[6] Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha."ZmBART: An Unsupervised Cross-lingual Transfer Framework for Language Generation." In Findings of the Association for Computational Linguistics: ACL 2021, pages 2804–2818, Online.

Auxiliary Task: Adaptive Pre-training

> "Given an input passage, generate few random sentences (called rand-summary) from the passage"

Data preparation steps for the auxiliary task are given below:

1. Generate a random number $k \in \{5. \cdots, 25\}$. k denotes the size of input passage
2. PASSAGE: Append k continuous sentences, starting from a random index of monolingual corpus D_i of the i^{th} language
3. RAND-SUMMARY: Randomly select 20% sentences from the passage
4. Repeat steps 1 to 3 for p languages
5. Repeat steps 1 to 4 for N times, to collect Np <PASSAGE, RAND-SUMMARY> pairs



[6] Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha."ZmBART: An Unsupervised Cross-lingual Transfer Framework for Language Generation." In Findings of the Association for Computational Linguistics: ACL 2021, pages 2804–2818, Online.

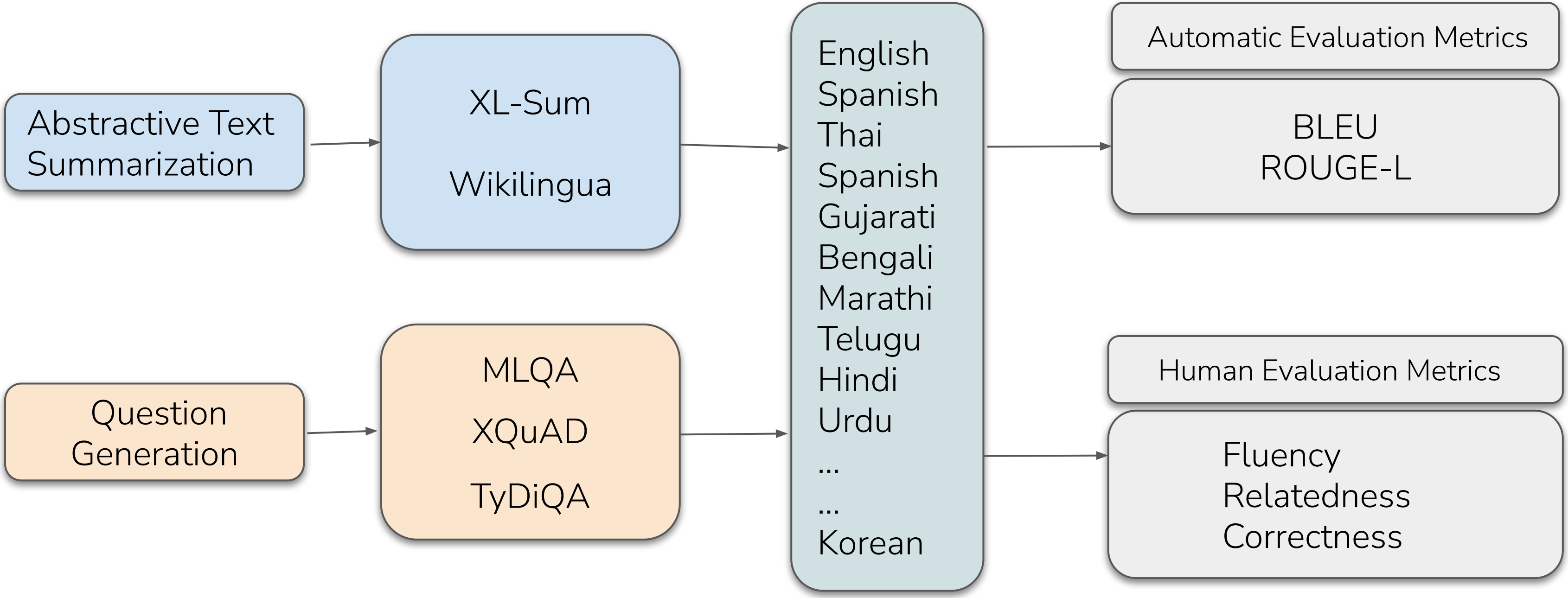
What Worked?

1. Added Language Tag: <fxx><2xx>
2. Adaptive Unsupervised Pre-training
3. Freezing model Components: Freeze token embedding and all the parameters of decoder layers

[6] Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha."ZmBART: An Unsupervised Cross-lingual Transfer Framework for Language Generation." In Findings of the Association for Computational Linguistics: ACL 2021, pages 2804–2818, Online.

❑ Meta-XNLG: Experimental Setup

Experimental Setup



Two Tasks | **Five Datasets** | **30 Languages** | **Five Evaluation Metrics**

Baselines

- **EnZmT5:** Inspired from [6]. Adaptive pre-training + supervised fine-tuning with HRL + zero-shot evaluation in LRLs

- **FTZmT5:** Inspired from [21]. Fine-tune EnZmT5 with centroid Languages + zero-shot evaluation in LRLs

[6] Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha."ZmBART: An Unsupervised Cross-lingual Transfer Framework for Language Generation." In Findings of the Association for Computational Linguistics: ACL 2021.

[21] Lewis, M., Ghazvininejad, M., Ghosh, G., Aghajanyan, A., Wang, S., & Zettlemoyer, L. (2020). Pre-training via paraphrasing. Advances in Neural Information Processing Systems, 33, 18470-18481.

❑ Meta-XNLG: Results and Analyses

Results: Automated Evaluation

XI-SUM

Model	fr	gu	id	th	ta	hi	mr	ja	ko	tr	ru	sw	pt	ar	te	ur	ne	bn	zh
EnZmT5	18.45	13.21	19.77	21.53	11.58	22.24	11.89	22.81	18.74	17.72	15.27	18.91	18.92	18.44	10.77	21.61	**16.24**	16.12	21.07
FTZmT5	21.83	7.98	19.27	**24.68**	10.80	11.92	8.94	23.32	16.82	14.99	12.90	21.01	20.07	15.85	9.14	13.05	11.06	12.66	15.20
Meta-X_{NLG}	**22.83**	**14.02**	**21.54**	24.61	**12.88**	**23.09**	**12.58**	**25.33**	**20.12**	**18.65**	**17.31**	**22.63**	**20.24**	**20.11**	**12.07**	**23.41**	15.45	**17.96**	**22.95**

Wikilingua

Model	id	fr	ar	pt	it	th	ru	cs	nl	de	ja	zh	hi	tr
EnZmT5	15.34	18.72	**15.70**	17.21	15.05	26.66	14.67	9.42	17.97	13.69	22.32	20.12	18.88	14.45
FTZmT5	13.69	19.37	12.66	17.80	15.54	23.72	11.95	10.20	16.74	12.22	22.81	18.64	17.32	13.84
Meta-X_{NLG}	**16.85**	**20.26**	15.66	**18.36**	**16.03**	**27.71**	**14.89**	**11.76**	**19.09**	**14.11**	**22.83**	**22.45**	**19.60**	**15.23**

MLQA

Model	ar	de	zh	vi	hi	el	ru	ro
EnZmT5	8.55	9.99	23.76	17.29	9.55	8.18	10.98	11.27
FTZmT5	5.82	9.040	22.87	16.47	9.05	6.95	8.87	10.31
Meta-X_{NLG}	**8.63**	**10.52**	**24.89**	**20.92**	**11.90**	**9.01**	**11.41**	**12.24**

Model	fi	ru	id	sw	ko	bn	ta
EnZmT5	7.87	5.52	5.75	4.48	8.59	5.77	3.08
FTZmT5	8.39	7.28	**11.42**	5.51	10.05	7.96	2.022
Meta-X_{NLG}	**9.08**	**7.47**	9.36	**6.42**	**12.67**	**9.17**	**9.76**

XQuAD

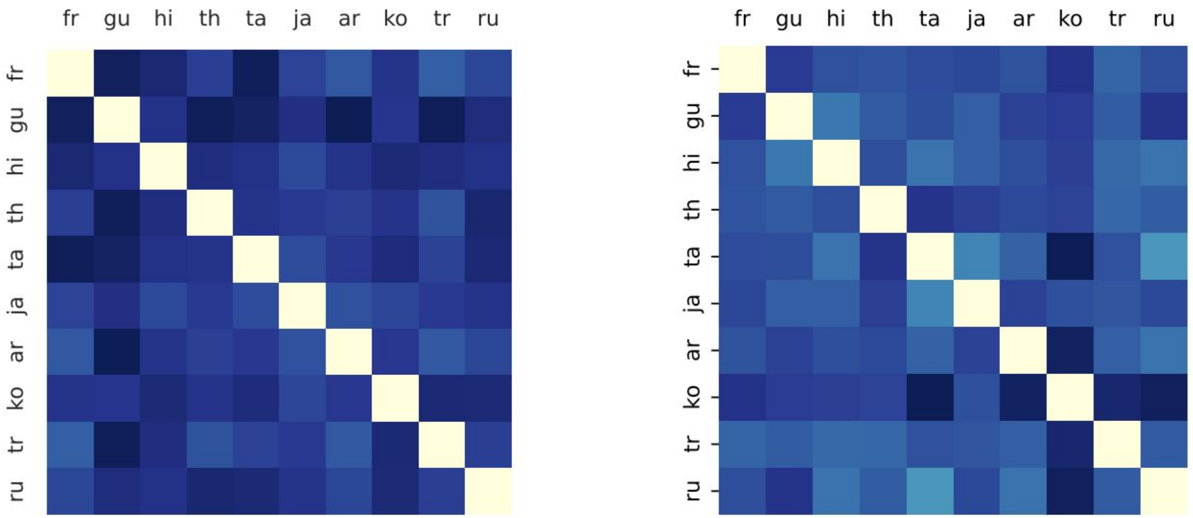
Model	hi	es	ar	zh
EnZmT5	5.06	6.94	3.46	13.70
FTZmT5	5.14	6.16	2.21	13.38
Meta-X_{NLG}	**5.66**	**7.03**	**3.66**	**15.13**

TyDiQA

Results: Human Evaluation

Model	**Task/Data/Lang**	**Flu**	**Rel**	**Corr**	**Task/Data/Lang**	**Flu**	**Rel**	**Corr**
Annotator set-1								
EnZmT5	ATS/XL-Sum/bn	4.06	3.58	2.84	ATS/XL-Sum/te	4.28	3.94	3.70
FTZmT5		2.82	3.18	2.08		3.46	3.46	3.22
Meta-X$_{NLG}$		**4.12**	**4.34**	**3.44**		**4.50**	**4.22**	**4.04**
Annotator set-2								
EnZmT5	ATS/XL-Sum/bn	3.70	3.23	3.26	ATS/XL-Sum/te	3.56	3.50	3.20
FTZmT5		2.62	2.48	2.16		3.02	2.84	2.60
Meta-X$_{NLG}$		**3.97**	**3.48**	**3.28**		**4.18**	**4.10**	**3.88**
Annotator set-1								
EnZmT5	ATS/Wiki/hi	4.00	3.72	3.68	QG/XQuAD/hi	4.12	4.24	2.54
FTZmT5		4.07	3.39	3.83		4.22	4.02	2.56
Meta-X$_{NLG}$		**4.09**	**3.80**	**3.97**		**4.42**	**4.34**	**2.86**
Annotator set-2								
EnZmT5	ATS/Wiki/hi	4.38	4.22	4.00	QG/XQuAD/hi	3.28	3.63	2.82
FTZmT5		4.57	**4.44**	4.08		3.24	3.34	2.89
Meta-X$_{NLG}$		**4.66**	**4.44**	**4.16**		**3.59**	**3.67**	**3.24**
Annotator set-1								
EnZmT5	QG/MLQA/hi	3.48	3.70	3.46	QG/TyDiQA/ta	4.25	4.06	3.10
FTZmT5		3.44	3.42	3.18		3.25	3.01	2.07
Meta-X$_{NLG}$		**3.70**	**3.74**	**3.56**		**4.74**	**4.20**	**3.39**
Annotator set-2								
EnZmT5	QG/MLQA/hi	**3.30**	3.28	2.40	QG/TyDiQA/ta	3.00	4.08	2.82
FTZmT5		3.10	3.44	2.84		2.55	3.045	1.83
Meta-X$_{NLG}$		3.24	**3.70**	**2.88**		**4.04**	**4.46**	**3.20**

Analysis-I: Cross-lingual Transfer



(a) Baseline (b) Meta-X_{NLG}

Cosine distance between language tags obtained from EnZmT5 and Meta-XNLG for 10 languages from XL-Sum dataset. Dark color indicate higher cosine distance.

Analysis-II: Effect of Meta-Training Languages

SetUp	MTrain Lang	ar	de	zh	vi	hi	el	ru	ro	avg
1	tr	6.14	8.61	23.67	19.81	**10.91**	6.80	9.53	10.17	11.89
2	es	6.68	10.82	20.89	16.84	7.96	7.79	10.02	13.28	11.78
3	th	5.43	8.47	23.10	17.46	7.99	6.85	9.41	8.98	11.08
4	ro	4.78	9.49	19.80	15.75	6.01	-	8.25	9.90	10.56
5	es,th	6.07	10.30	18.74	16.10	7.74	7.14	9.56	12.37	11.00
6	tr,th	6.02	8.58	25.05	19.08	10.38	6.64	9.27	10.40	11.92
7	ro,de	5.53	-	22.69	15.37	7.59	6.37	8.85	-	11.06
8	zh,ar	-	8.92	-	15.55	8.22	6.58	9.72	10.49	9.91
9	de,ru	6.02	-	17.68	12.40	8.05	7.32	-	12.56	10.67
10	vi,th, el	6.15	9.86	23.26	-	8.86	-	9.94	11.71	11.63
11	de,tr,el	5.91	-	14.29	18.15	9.50	-	9.88	12.28	11.66
12	tr,es,th, ru	6.03	**11.88**	23.13	19.56	9.58	7.04	-	**13.62**	12.97
13	tr,es,th,de	6.34	-	17.25	19.47	8.91	7.73	9.95	13.14	11.82
14	tr,es,th,de,ru	6.45	-	**25.14**	16.31	9.51	6.72	-	12.39	12.75
15	tr,es,th,de,ru,ar	-	-	22.58	15.65	8.04	6.74	-	11.81	12.96
16	Meta-X_{NLG}	**8.63**	10.52	24.89	**20.92**	**11.90**	**9.01**	**11.41**	12.24	**13.69**

- Zero-shot results on different training languages combinations of the XQuAD dataset.
- '-' indicates the language used in training, so scores are not zero-shot and not included.

❑ Meta-XNLG: Summary

Summary

- We propose a novel Meta-XNLG framework based on meta-learning and language clustering to uniformly transfer supervision.
- This is the first study that uses meta-learning for zero-shot cross-lingual transfer and generation.
- The evaluations are done on two challenging NLG tasks (ATS and QG), five publicly available datasets and 30 languages and consistent improvements are observed.

NATURAL LANGUAGE
& INFORMATION
PROCESSING LAB

EMNLP
2023

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

SelectNoise: Unsupervised Noise Injection to Enable Zero-Shot Machine Translation for Extremely Low-resource Languages

Maharaj Brahma* and **Kaushal Kumar Maurya*** and Maunendra Sankar Desarkar

Natural Language & Information Processing Lab (NLIP Lab)
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad, India

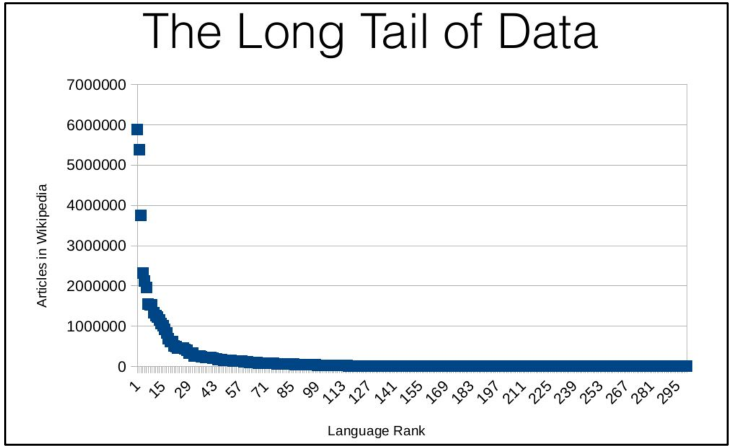
  

*Equal Contribution

- SELECTNOISE: Introduction and Motivation

Recollect: Language Landscape

- 7000+ languages across the globe [3]
- Around only 300 languages have wikipedia articles
- Languages data resources availability follows long-tail distribution
- Majority of research focus on English - Less Inclusivity and Diversity [3, 4]



Source: Graham Neubig Multilingual NLP Lectures

[3] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 2020 6282–6293.
[4] Bender, Emily. "The# benderrule: On naming the languages we study and why it matters." *The Gradient* 14 (2019).

Introduction: Extremely LRLs (ELRLs)

- Lack parallel data
- Limited monolingual data
- Representations are absent from existing multilingual pre-trained language models

SELECTNOISE: Problem Statement

"Machine Translation from ELRL to English in the zero-shot setting."

Literature Review: MT for LRLs

- Cross lingual transfer among languages: Multilingual NMT
- Reduce reliance of parallel data: Unsupervised NMT
- Monolingual corpus incorporated NMT: Back-translation
- Data augmentation approaches for MT: Perturbation Models
- BPE vocabulary overlapping among related languages [22]

Limited Efforts has been made for ELRL for MT task

[22] Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland. Association for Computational Linguistics.

Motivation: Hopeful direction

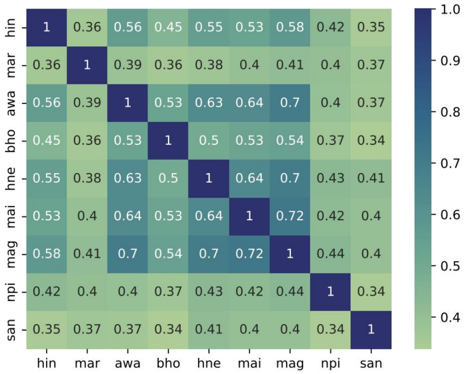
- Utilize relatedness among languages
  - Dialectal variations
  - Vocabulary sharing
  - Similarities due to Geographical proximity

- Many ELRLs are related with some High resource Language (HRL)

Hindi: कनाडियन के खिलाफ नडाल का सीधा रिकॉर्ड 7-2 है।

Bhojpuri: कनाडा के खिलाफ़ नाडाल के हेड-टू -हेड रिकॉर्ड 7-2 के बा।

Lexical level similarity between languages

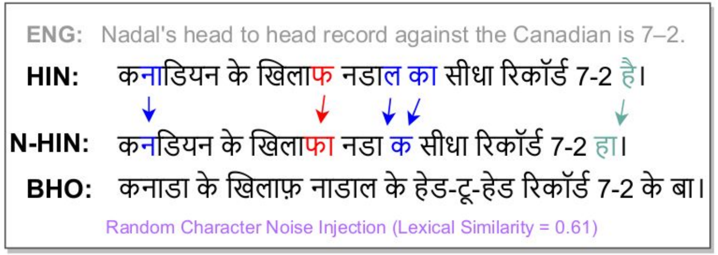


Lexical Similarity heatmap

Motivation: Hopeful direction

Earlier Success for ELRL:

- Recall: Exploit lexical similarity through random noise augmentation [23]



Limitations:

- Studies limited to NLU tasks only
- Random Noise Injection in HRL may be suboptimal for NLG task especially MT as injections are random
- Noising strategy should be systematic and incorporate linguistic signals

[23] Noëmi Aepli and Rico Sennrich. 2022. Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

❑ SELECTNOISE: Proposed Approach

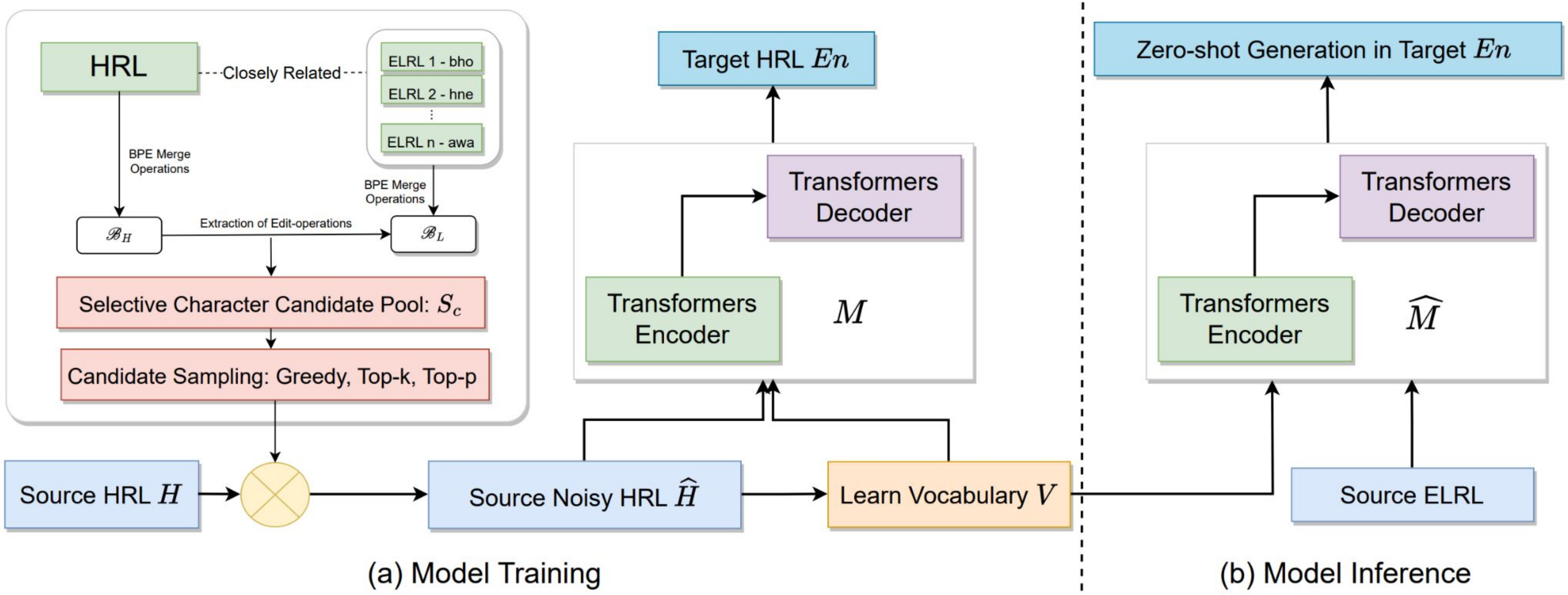
SelectNoise: Overview

- Modeling Approach: Introduce selective character noise augmentation to improved cross-lingual transfer from HRL to ELRLs.
- Noise Augmentation: In the source side of HRL to English parallel data.
- Proxy Training Data: Utilize noise-augmented parallel data as proxy training data for ELRL to English translation task.
- Noise Candidate Extraction: Extract noise injection candidates using BPE merge operations and edit operations, known as *selective Candidate*.
- Sampling Algorithms: Greedy, top-k, and top-p.

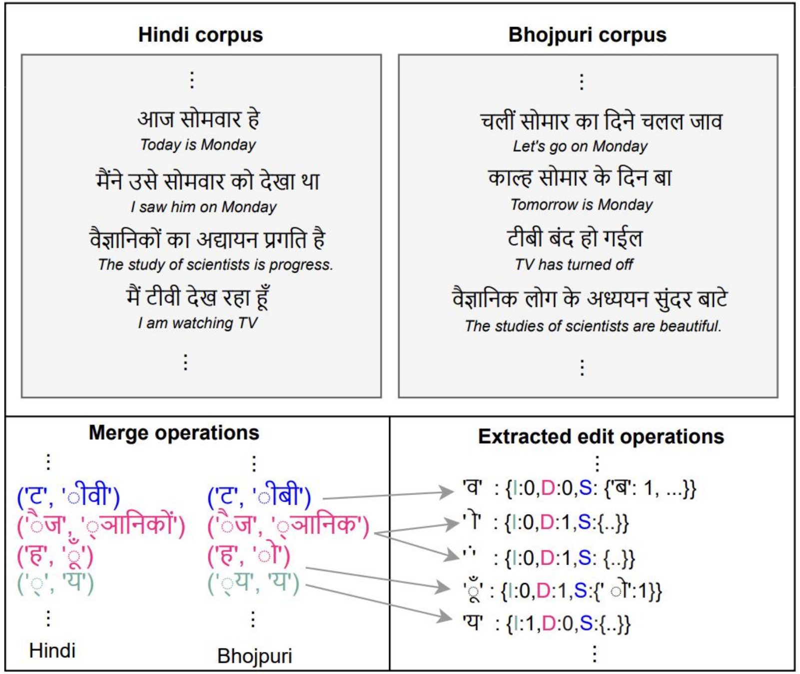
SelectNoise: Overview

- **Intuition:**
  - Noise augmentation act as regularizer.
  - Facilitates better a cross-lingual transfer
- **Hypothesis:**
  - SelectNoise is expected to outperform random noise augmentation approaches.
  - SelectNoise performance should match supervised noise augmentation model.

SELECTNOISE: Overview



SELECTNOISE: Candidate Extraction



BPE Merge operation and Edit-operations

Candidate Pool (S_c) Template

{
C_1: {'I': f_1, 'D': f_3, 'S': { E_1: f_4, E_2: f_5, ... }},
C_2: {'I': f_4, 'D': f_3, 'S': { E_3: f_6, E_2: f_7,... }},
⋮
C_i: {'I': f_2, 'D': f_4, 'S': { E_1: f_2, E_3: f_1, ... }}
⋮
C_n: {'I': f_1, 'D': f_4, 'S': { E_1: f_4, E3: f_5, ... }}
}

Few Sample Elements of Candidate Pool (S_c)

```
{
 '्':{'I':62,'D':1561,'S':{'ह': 1482,...}}
 'र':{'I':92,'D':97,'S':{'ा':1482,...}}
 'ि':{'I':1552,'D':15,'S':{'य':397,...}}
 'S': {'I': 0,'D': 33, 'S': {}}
}
```

Selective Character Candidate Pool

SelectNoise: Candidate Augmentation

Candidate Pool (S_c) Template

{

C_1: {'I': f_1, 'D': f_3, 'S': { E_1: f_4, E_2: f_5, ... }},
C_2: {'I': f_4, 'D': f_3, 'S': { E_3: f_6, E_2: f_7,... }},

⋮

C_i: {'I': f_2, 'D': f_4, 'S': { E_1: f_2, E_3: f_1, ... }}

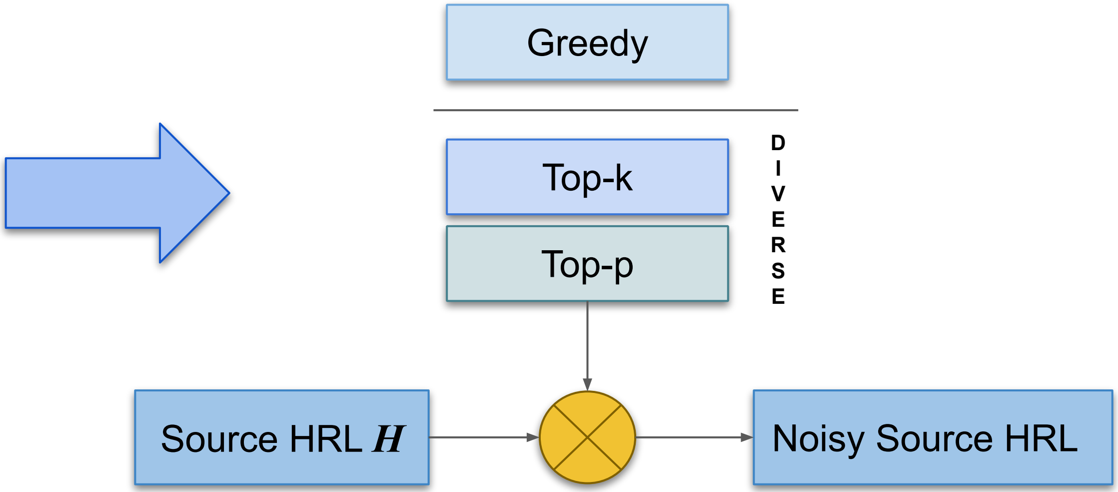
⋮

C_n: {'I': f_1, 'D': f_4, 'S': { E_1: f_4, E3: f_5, ... }}

}

Few Sample Elements of Candidate Pool (S_c)

```
{
 '्':{'I':62,'D':1561,'S':{'ह': 1482,...}}
 'र':{'I':92,'D':97,'S':{'ा':1482,...}}
 'ि':{'I':1552,'D':15,'S':{'य':397,...}}
 'ऽ': {'I': 0,'D': 33, 'S': {}}
}
```



SelectNoise: Training and Evaluation Setups

- Constraints: HRLs and LRLs should be closely related
- Data Sources:
  - No parallel data for ELRLs.
  - 1000 monolingual examples for each ELRLs
- Model Training: No pre-trained LLMs, trained from scratch.
- Operations [char-level]: Insertion, deletion and substitution.
- Noise Injection Percentage: Injected noise at 5-10%.
- Zero-shot Evaluation:
  - Trained on proxy HRL parallel data.
  - Evaluated with unseen ELRLs

SelectNoise: Why it should work?

ENG: Nadal's head to head record against the Canadian is 7–2.

HIN: कनाडियन के खिलाफ नडाल का सीधा रिकॉर्ड 7-2 है।

N-HIN: कनडियन के खिलाफा नडा क सीधा रिकॉर्ड 7-2 हा।

BHO: कनाडा के खिलाफ़ नाडाल के हेड-टू-हेड रिकॉर्ड 7-2 के बा।

Random Character Noise Injection (Lexical Similarity = 0.61)

ENG: Nadal's head to head record against the Canadian is 7–2.

HIN: कनाडियन के खिलाफ नडाल का सीधा रिकॉर्ड 7-2 है।

N-HIN: कनडियन के खिलाफ़ नाडाल के सीधा रिकॉर्ड 7-2 बा।

BHO: कनाडा के खिलाफ़ नाडाल के हेड-टू-हेड रिकॉर्ड 7-2 के बा।

SelectNoise Model (Lexical Similarity = 0.77)

❏ SELECTNOISE: Experimental Setup

Experimental Setup

2 Language Families

Indo-Aryan

- Hindi (HRL) — AI4Bharat [9]
- Bhojpuri (bho)
- Chhattisgarhi (hne)
- Sanskrit (san)
- Maithili (mai)
- Magahi (mag)
- Awadhi (awa)
- Nepali (npi)
- Kashmiri (kas)

FLORES 200 (ELRLs)

Romance

- ES (HRL) — Rapp (2021) [10]
- Catalan (cat)
- Galician (glg)
- Asturian (ast)
- Occitan (oci)

FLORES 200 (ELRLs)

Automatic evaluation

- BLEU
- chrF
- BLEURT
- COMET

Human evaluation

- XSTS [30]

7 Baselines

- Vanilla NMT [24]
- Word-drop [26]
- BPE-drop[25]
- SwitchOut [27]
- OBPE [29]
- BPE-Dropout [28]
- Random Char Noise [23]

2 Language Families | 12 ELRLs; 2 HRLs | Datasets | Evaluation Metrics | 7 Baselines

❑ SelectNoise: Results and Analyses

Automated Evaluation Results: ChrF Scores

Models	Indo-Aryan								Romance				Average
	bho	hne	san	mai	mag	awa	npi	kas	cat	glg	ast	oci	
Vanilla NMT	40.3	46.8	22.3	40.0	49.3	47.6	29.6	21.3	33.0	41.0	40.7	33.0	37.08
Word-drop	39.5	47.2	21.8	40.6	49.0	47.6	28.6	20.6	37.6	43.6	43.4	36.0	37.96
BPE-drop	39.1	46.8	22.6	40.4	48.7	46.7	29.2	21.1	33.8	41.7	41.5	33.0	37.05
SwitchOut	36.1	43.2	20.1	38.2	45.6	42.7	28.3	18.8	29.0	34.9	34.9	29.1	33.41
OBPE	41.3	47.5	23.4	41.8	50.4	49.7	30.5	21.1	34.1	41.2	41.3	33.8	38.00
BPE-Dropout	39.8	47.4	22.5	39.9	49.6	47.7	29.3	21.2	33.2	40.8	41.4	33.0	37.15
Random Char Noise	40.9	48.4	23.8	40.8	50.0	47.5	31.2	21.9	40.9	46.1	46.4	38.2	39.68
SELECTNOISE Model													
SELECTNOISE + Greedy	42.1	**51.0**	25.2	**43.4**	**51.7**	**49.9**	33.4	**23.7**	**42.0**	**47.1**	47.4	38.5	**41.28**
SELECTNOISE + Top-k	**42.4**	49.9	**26.0**	43.0	51.0	48.8	33.4	23.3	41.5	47.1	**47.8**	38.5	41.06
SELECTNOISE + Top-p	42.0	49.6	24.1	42.4	50.6	48.8	**33.6**	23.3	41.6	47.1	47.5	**38.8**	40.78
Supervised Noise Injection Model													
Selective noise + Greedy	41.4	49.1	25.4	42.2	50.1	48.7	32.9	22.2	41.6	47.2	47.7	38.7	40.60
Selective noise + Top-k	41.7	49.3	26.3	43.3	50.8	48.7	34.2	23.6	41.9	46.8	47.5	38.7	41.10
Selective noise + Top-p	41.4	49.9	27.3	43.3	51.6	48.9	33.9	23.4	41.6	47.7	48.2	39.0	41.35

Zero-shot chrF scores for ELRLs → English

- Similar improvements for BLEU, COMET and BLEURT metrics

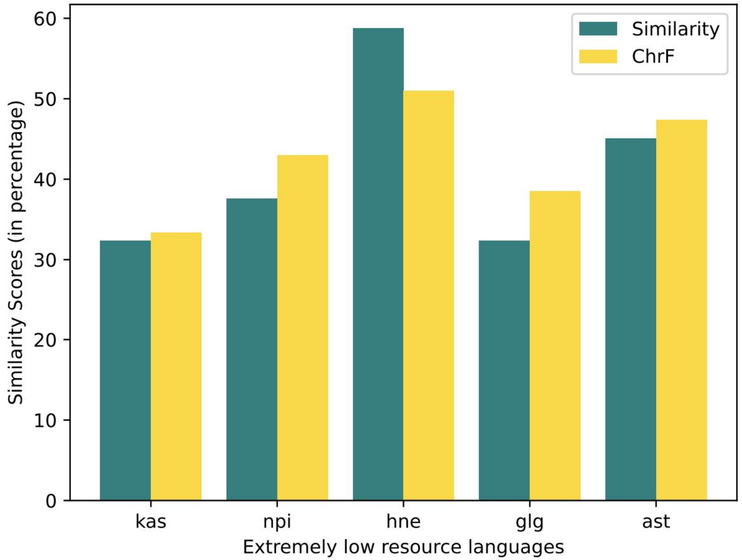
Human Evaluation Results: XSTS Scores

Models	Languages		
	bho	**san**	**npi**
Annotator set-1			
Vanilla NMT	3.54	2.42	2.21
BPE Dropout	3.29	2.37	1.83
SELECTNOISE Model	**4.17**	**2.83**	**2.50**
Annotator set-2			
Vanilla NMT	3.42	1.96	2.17
BPE Dropout	2.79	1.83	1.96
SELECTNOISE Model	**3.54**	**2.17**	**2.21**

- Evaluation on 24 examples for each language
- Cross Lingual Semantic Text Similarity (XSTS)[30] metric scores between 1-5

[30] AI, Daniel Licht META, Cynthia Gao META AI, Janice Lam META AI, Francisco Guzmán META AI, Mona Diab, and Philipp Koehn. "Consistent Human Evaluation of Machine Translation across Language Pairs." *Volume 1: MT Research Track* (2022).

Analysis: Language Similarity vs. Performance



Observation: The ELRL(s) that exhibit higher lexical similarity with HRL lead to better performance.

Analysis: Impact of Monolingual Data Size

Language	Data size	BLEU	chrF
hne	997	19.5	49.6
	6000	**20.3**	**50.3**
mai	997	11.9	42.4
	6000	**12.4**	**43.2**
npi	997	6.7	33.6
	6000	**7.2**	**33.8**

Observation: Larger the monolingual data, more accurate extraction of selectively candidates, resulting in improved performance.

❑ SELECTNOISE: Summary

SelectNoise: Summary

- SelelctNoise model outperforms strong baselines across 12 ELRLs for ELRLs → English MT task
- The model requires no parallel data, only limited monolingual data, and no LLM multilingual representation
- Unsupervised noise injection gives comparable performance with Supervised approach
- Highly Scalable
- Cumulative gain of 11.3% chrF over Vanilla-NMT

Thesis Conclusion

- With this thesis we made an effort to extend neural text generation in three diverse dimensions
  - Diverse text generation
  - Text generation with limited context
  - Text generation with limited data/supervision for LRLs

- We have developed the novel modeling approaches the tackle each of problems and sub-problems

- With the introduction of large language models - we still believe that- the proposed modeling approaches will hold value, especially in low-resource settings, as the large language models are not inherently multilingual and not scalable to LRLs (or ELRLs).

Future Research Directions

- Unified Modeling for diverse text generation
- Advancing the RAG modeling
- Language technology for next 7000+ languages
- Modeling towards multilinguality
- Evaluation of multilingual NLG
- Evaluation without reference
- Many more..

Publications Included in the Thesis

1. **K. K. Maurya**, and M. S. Desarkar. "Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors for multiple-choice questions for reading comprehension." in **CIKM 2020**.
2. **K. K. Maurya**, M. S. Desarkar, Y. Kano, and K. Deepshikha."ZmBART: An Unsupervised Cross-lingual Transfer Framework for Language Generation." In Findings of **ACL 2021.**
3. **K. K. Maurya** and M. S. Desarkar. "Meta-XNLG: A Meta-Learning Approach Based on Language Clustering for Zero-Shot Cross-Lingual Transfer and Generation." In Findings of **ACL 2022**.
4. **K. K. Maurya,** M. S. Desarkar, M. Gupta, and P. Agarwal. "Trie-NLG: trie context augmentation to improve personalized query auto-completion for short and unseen prefixes." in **ECML-PKDD (DAMI) 2023**.
5. M. Brahma*, **K. K. Maurya***, and M. S. Desarkar. "SelectNoise: Unsupervised Noise Injection to Enable Zero-Shot Machine Translation for Extremely Low-Resource Languages." In Findings of **EMNLP 2023**.
6. **K. K. Maurya**, R. Kejriwal, M. S. Desarkar and A. Kunchukuttan. "CharSpan: Utilizing Lexical Similarity to Enable Zero-Shot Machine Translation for Extremely Low-Resource Languages." In **EACL 2024**.

* Equal Contribution

Publications NOT Included in the Thesis

1. Sreekanth Madisetty, **Kaushal Kumar Maurya,** Akiko Aizawa, and Maunendra Sankar Desarkar. “A neural approach for detecting inline mathematical expressions from scientific documents.” **Expert Systems** 38, no. 4 (2021): e12576.

2. Arkadipta De, Venkatesh E, **Kaushal Kumar Maurya**, Maunendra Sankar Desarkar, “Coarse and FineGrained Hostility Detection in Hindi Posts using Fine Tuned Multilingual Embeddings.” CONSTRAINT workshop, **AAAI-w 2021**. **Shared task best paper honorable mention.**

3. Aditi Bagora, Kamal Shrestha, **Kaushal Kumar Maurya**, and Maunendra Sankar Desarkar. “Hostility Detection in Online Hindi-English Code-Mixed Conversations.” In Proceedings of the 14th ACM Web Science Conference 2022, pp. 390-400. 2022. **WebSci 2022**.

4. Venkatesh, E., **Kaushal Kumar Maurya**, Deepak Kumar, and Maunendra Sankar Desarkar. “DivHSK: Diverse Headline Generation using Self-Attention based Keyword Selection.” In Findings of the Association for Computational Linguistics: **ACL 2023**, pp. 1879-1891. 2023.

Publications NOT Included in the Thesis [Cont..]

5. **Kaushal Kumar Maurya** and Maunendra Desarkar. "Towards Low-resource Language Generation with Limited Supervision." In the BigPicture workshop. Empirical Methods in Natural Language Processing: **EMNLP-w 2023**, Singapore.

6. Aishwarya M*, **Kaushal Kumar Maurya***, Manish Gupta and Maunendra Desarkar. "DQAC: Detoxifying Query Auto-Completion with Adapters." In **PAKDD 2024**, Taipei, Taiwan

7. Aishwarya M, **Kaushal Kumar Maurya**, Manish Gupta and Maunendra Desarkar. "DAC: Quantized Optimal Transport Reward-based Reinforcement Learning Approach to Detoxify Query Auto-Completion." 2023. Under Review.

8. **Kaushal Kumar Maurya**, Aishwarya M, Manish Gupta and Maunendra Desarkar. "ECoAdapters: Efficient Multilingual Controlled Text Generation with Adapters." 2023. Under Review.

* Equal Contribution

Acknowledgements

- PhD Supervisor: Dr. Maunendra Sankar Desarkar
- Doctoral committee members: Profs. Vineeth N. Balasubramanian, Dr. Srijith P.K., and Prof. J. Balasubramaniam
- All the anonymous reviewers, meta-reviewers, human evaluators and proofreaders
- Collaborating Organizations: Microsoft, Nvidia, Shizuoka University, and the University of Tokyo.
- All collaborators, friends, colleagues and family members
- Many more (list is very large)

References-I

[1] Maurya, Kaushal Kumar, and Desarkar, Maunendra Sankar. "Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors for multiple-choice questions for reading comprehension." *Proceedings of the 29th ACM international conference on information & knowledge management*. 2020.

[2] Maurya, Kaushal Kumar, Maunendra Sankar Desarkar, Manish Gupta, and Puneet Agrawal. "trie-nlg: trie context augmentation to improve personalized query auto-completion for short and unseen prefixes." *Data Mining and Knowledge Discovery* 37, no. 6 (2023): 2306-2329.

[3] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 2020 6282–6293.

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

[5] Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4232–4267, Singapore. Association for Computational Linguistics.

[6] Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha."ZmBART: An Unsupervised Cross-lingual Transfer Framework for Language Generation." In Findings of the Association for Computational Linguistics: ACL 2021, pages 2804–2818, Online.

[7] Kaushal Kumar Maurya and Maunendra Desarkar. "Meta-XNLG: A Meta-Learning Approach Based on Language Clustering for Zero-Shot Cross-Lingual Transfer and Generation." In Findings of the Association for Computational Linguistics: ACL 2022, pages 269–284, Dublin, Ireland.

[8] Maharaj Brahma Kaushal Kumar Maurya 3 , and Maunendra Desarkar. "SelectNoise: Unsupervised Noise Injection to Enable Zero-Shot Machine Translation for Extremely Low-Resource Languages." In Findings of the Empirical Methods in Natural Language Processing: EMNLP 2023, Singapore.

[9] Kaushal Kumar Maurya, Rahul Kejriwal, Maunendra Desarkar and Anoop Kunchukuttan. "CharSpan: Utilizing Lexical Similarity to Enable Zero-Shot Machine Translation for Extremely Low-Resource Languages." In The 18th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2024.

[10] Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. Overcoming catastrophic forgetting in zero-shot cross- lingual generation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9279–9300, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

References-II

[11] Liang, Yaobo, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong et al. "XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation." *arXiv preprint arXiv:2004.01401* (2020).

[12] Hu, Junjie, et al. "Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation." International Conference on Machine Learning. PMLR, 2020.

[13] Guokun Lai, Barlas Oguz, Yiming Yang, and Veselin Stoyanov. 2019. Bridging the domain gap in cross-lingual document classification. CoRR, abs/1909.07009.

[14] Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

[15] Bengio, Yoshua, Samy Bengio, and Jocelyn Cloutier. Learning a synaptic learning rule. Université de Montréal, Département d'informatique et de recherche opérationnelle, 1990.

[16] Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." International conference on machine learning. PMLR, 2017.

[17] Dryer, Matthew S., and Martin Haspelmath. "The world atlas of language structures online." (2013).

[18] Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics

[19] Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2391–2406, Online. Association for Computational Linguistics.

[20] 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, pages 483–498, Online. Association for Computational Linguistics.

[21] Lewis, M., Ghazvininejad, M., Ghosh, G., Aghajanyan, A., Wang, S., & Zettlemoyer, L. (2020). Pre-training via paraphrasing. Advances in Neural Information Processing Systems, 33, 18470-18481.

[22] Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland. Association for Computational Linguistics.

References-III

[23] Noëmi Aepli and Rico Sennrich. 2022. Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

[24] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

[25] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

[26] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

[27] Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 856–861, Brussels, Belgium. Association for Computational Linguistics.

[28] Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1882–1892, Online. Association for Computational Linguistics.

[29] Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland. Association for Computational Linguistics.

[30] AI, Daniel Licht META, Cynthia Gao META AI, Janice Lam META AI, Francisco Guzmán META AI, Mona Diab, and Philipp Koehn. "Consistent Human Evaluation of Machine Translation across Language Pairs." *Volume 1: MT Research Track* (2022).

NATURAL LANGUAGE
& INFORMATION
PROCESSING LAB

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Thank You!!!

NLIP Lab

Homepage

Contact us:
cs18resch11003@iith.ac.in
nlip@cse.iith.ac.in
https://nlip-lab.github.io/
https://kaushal0494.github.io/