ZmBART: An Unsupervised Cross-lingual Transfer Framework for Language Generation

[Findings of ACL 2021]

Kaushal Kumar Maurya, Maunendra Sankar Desarkar,
Yoshinobu Kano and Kumari Deepshikha
Department of CSE, IITH, Shizuoka University, Japan & NVAITC

Problem Statement

- Transfer *supervision* from High Resource Language (HRL) to Low Resource Languages (LRL) for natural language generation (NLG) without *Back-Translation* and without *Parallel/Pseudo-parallel Data*.

- Proposed unsupervised *framework (ZmBART)* to enable Zero/Few- shot language generation from mBART pre-trained Model.

Literature Review: Based on Machine Translation Pipeline

- Back Translation [2]
  - Language A->English->NLG model with English->Language A [2]
- Pseudo-training data [3,4]
  - (x_Eng, y_Eng) -> (x_Lang ,y_Lang)
- Incorporating machine translation and monolingual summarization dataset to improve the cross-lingual summarization. [5]

Limitations:

- MT systems are not perfect, so generated texts are error-prone
- These models are not scalable to low-resource languages as they don't share the latent space (representations) across languages.

Literature Review: Based on Parallel Data and Back-Translation

- [6] used a small annotated question generation (QG) dataset with back-translation for low-resource languages for cross-lingual QG.
- [7] proposed a pre-trained cross-lingual NLG (XNLG) model with parallel data for zero-shot transfer learning for question generation and text summarization.

Limitations:

- Parallel data is not available for all possible language pair
- Back-translation required Machine Translation systems
- Large task specific-data is not available for many languages

Proposed Approach: Goal

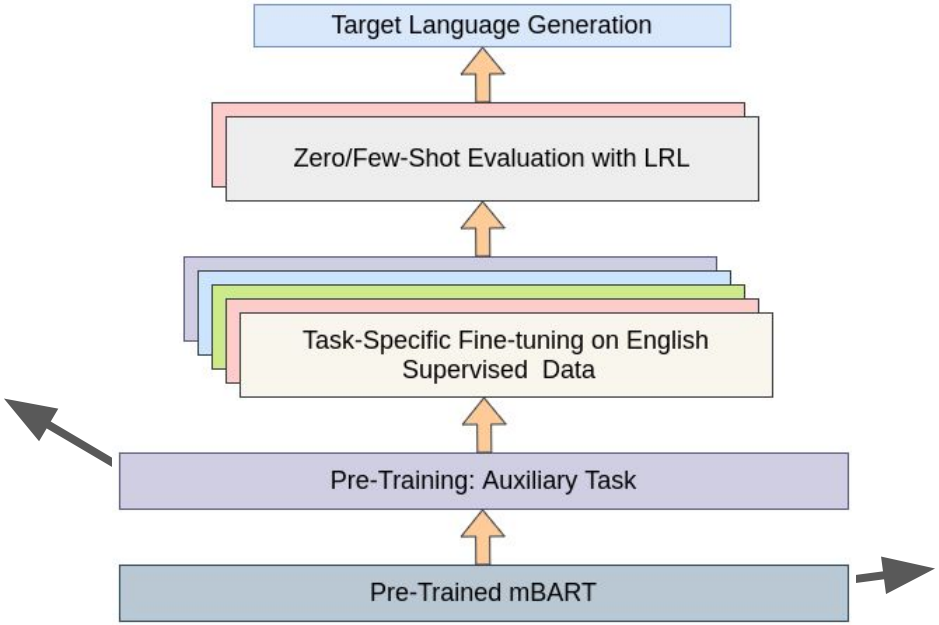
- Single Framework across multiple generation tasks (without even modifications in the hyperparameters)
- Only utilize monolingual data from three languages
- Leverage existing pre-trained multilingual models
- Framework should be simple and easy implementable
- Framework should be easily scalable

Proposed Approach: Areas in Focus

#Examples	Tasks	Languages
Zero shot Few Shot (100-1000)	Question Generation Distractor Generation Text Summarization News Headline Generation	English Hindi Japanese

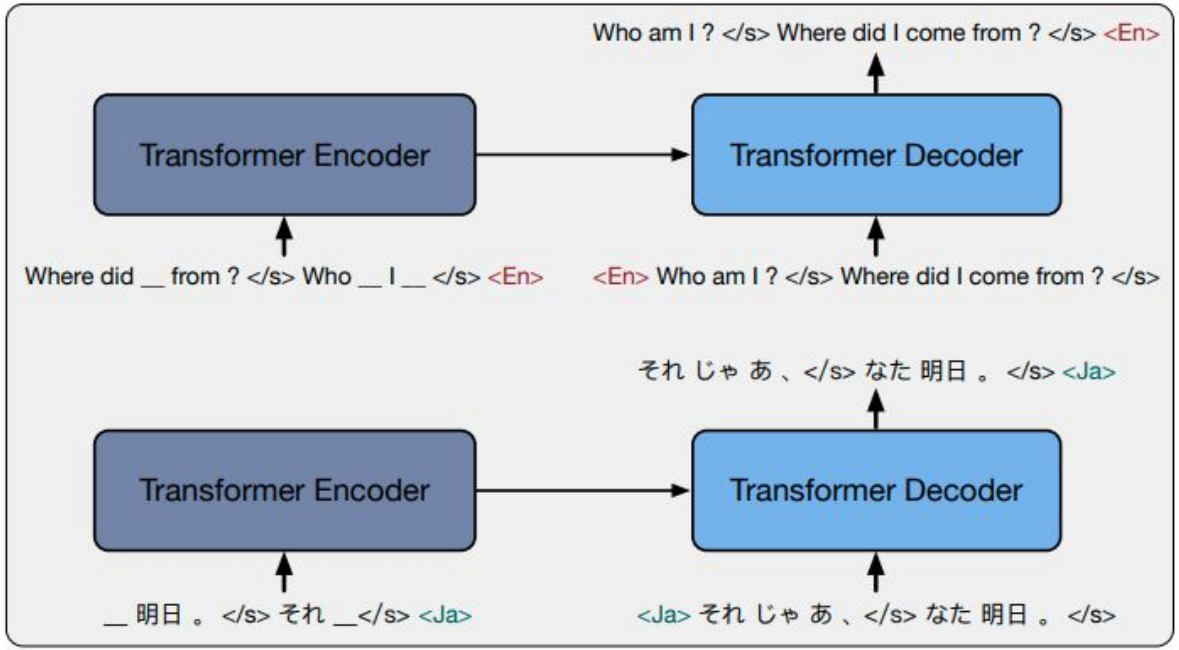
Proposed Approach: Architectural Diagram

Enriches the shared representations for selected languages



Provides latent representation space shared by multiple languages

mBART Model



Multilingual Denoising Pre-Training (mBART)

Source:Liu, Yinhan, et al. "Multilingual denoising pre-training for neural machine translation." *Transactions of the Association for Computational Linguistics* 8 (2020): 726-742.

Auxiliary Task: Motivation

mBART pre-trained model was trained with sentence shuffling and masking denoising objectives which does not directly follow auto-regressive decoding thereby causing mismatch between pre-training and fine-tuning objectives.

Auxiliary Task: Goal

- Additional pre-training step (aka: adoptive pre-training) for better warm-start to downstream
- Should only utilize monolingual data from selected languages
- Should enrich the mBART latent representation for selected languages
- Train the decoder in pure auto-regressive manner with a training objective which is close to multiple fine-tuning tasks.

Auxiliary Task: “further Pre-training” of mBART

“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. Randomly generate a number $k \in \{5\text{-}25\}$. 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

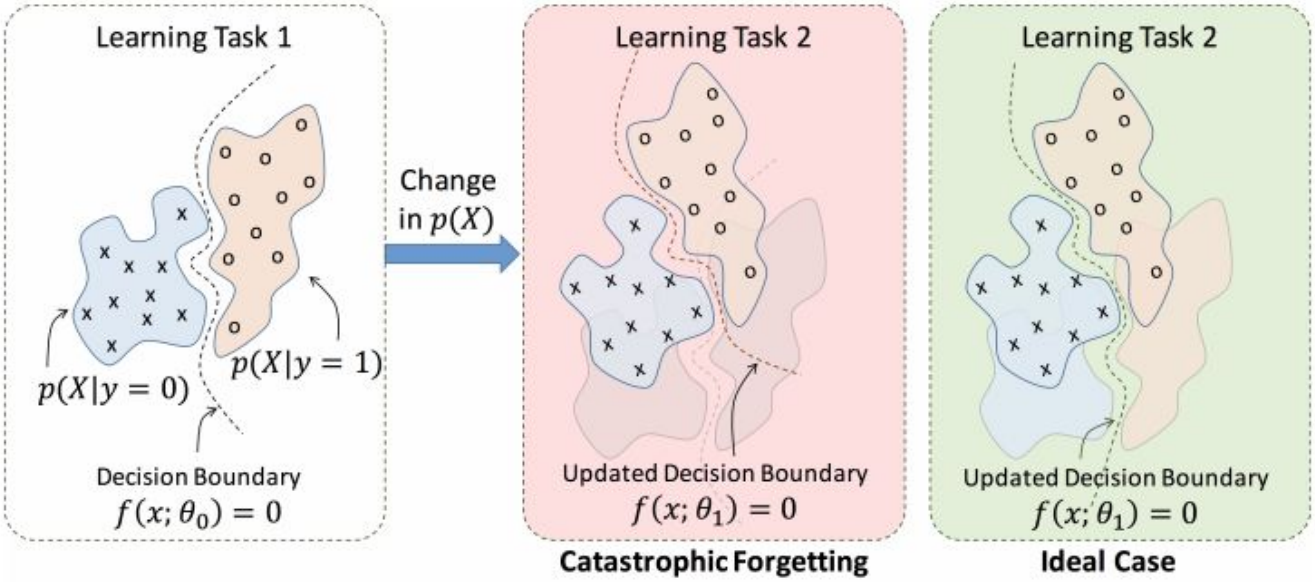
Sample Model Output

News Passage: दक्षिण कश्मीर के पुलवामा जिले में सुरक्षा बलों के साथ जारी मुठभेड़ में शुक्रवार को एक आतंकवादी ढेर हो गया.पुलिस के एक प्रवक्ता ने बताया कि इस मुठभेड़ में एक आतंकवादी मारा गया है. यह मुठभेड़ अभी जारी है.प्रवक्ता ने बताया कि पुलवामा के चन्दगाम में आज सुबह सुरक्षा बलों और छिपे हुए आतंकवादियों के बीच मुठभेड़ शुरू हो गई।माना जा रहा है कि गांव में लश्कर-ए-तैयबा के दो आतंकवादी छिपे हुए हैं।
(Translation: A militant was killed on Friday in an ongoing encounter with security forces in Pulwama district of eroded Kashmir. A police spokesman said a militant was killed in the encounter. The encounter is still going on, the spokesperson said, adding that an encounter between security forces and hidden militants started this morning at Chandgam in Pulwama. Two LeT militants are believed to be hiding in the village.)

Headline (ground truth): कश्मीर के पुलवामा में मुठभेड़, एक आतंकी ढेर
(Translation: Encounter in Pulwama, Kashmir, a terrorist killed)

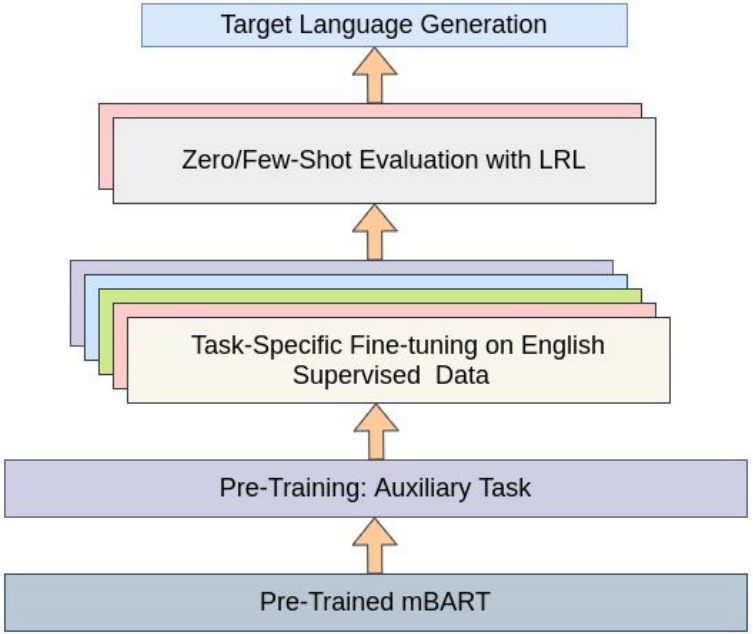
Headline (zero-shot generated output:) पुलवामा में जारी मुठभेड़ में एक आतंकवादी ढेर
(Translation: A terrorist killed in ongoing encounter in Pulwama)

Catastrophic Forgetting Problem



Source: Kolouri, Soheil, et al. "Attention-based structural-plasticity." *arXiv preprint arXiv:1903.06070* (2019).

Catastrophic Forgetting Problem



Fine-tuned model doesn't understand Hindi and japanese language and decoder only generate English text (also known as Accidental Translation[1]).

Source: Xue, Linting, et al. "mt5: A massively multilingual pre-trained text-to-text transformer." *arXiv preprint arXiv:2010.11934* (2020).

Solutions for Catastrophic Forgetting Problem

1. Freeze component(s) of the model
2. EWC: Penalize the loss function
3. Joint Training: Augmentation of few previous task data sample during the training of current task data

Freezing Model Components

Multiple choices:

- Word embeddings
- All encoder layers
- All decoder layers
- Subset of encoder layers
- Subset of decoder layers

Penalize the Loss Function: EWC Approach

$$L_{EWC}(\theta) = L^{S}_{NLL}(\theta) + \lambda \sum_{i} \bar{F}_{i,i} \left[\theta_i - \hat{\theta}^{G}_{i}\right]^2$$

Source: Thompson, Brian, et al. "Overcoming catastrophic forgetting during domain adaptation of neural machine translation." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019.

What worked for Us

Freezing Model Component: Freezing Word embedding and weights of all layers of the decoder

Framework Setting: Summary

Step-01: Adoptive training of mBART with Mono-lingual data (with Auxiliary task objective)
Step-02: Fine Tuning of task-01 Model with task-specific English data
Step-03: Zero-shot evaluation on task-02 model with Hindi and Japanese data

Catastrophic forgetting Problem: while fine-tuning with task-specific english data we freeze model component.

Source: Xue, Linting, et al. "mt5: A massively multilingual pre-trained text-to-text transformer." *arXiv preprint arXiv:2010.11934* (2020).

Tasks

Task	Definition
News Headline Generation (NHG)	Given a **news article**, we generate grammatically coherent, semantically correct and abstractive **headline**.
Question Generation (QG)	Given an **input passage and an answer**, the aim is to generate semantically and syntactically correct **question** that can produce the answer.
Abstractive Text Summarization (ATS)	In ATS, aim is to generate grammatically coherent, semantically correct and abstractive **summary** given an **input document**
Distractor Generation (DG)	For given **passage, question and answer** triplet, generate a long, coherent, and grammatically correct **wrong option**.

Dataset

Task	English Data Source	Hindi Data Source	Japanese Data Source
NHG	Gigaword New Headline (500k/30k/30k)	Kaggle (1k/1k/5k)	Shizuoka University (1k/1k/5k)
QG	SQuAD1.1 (80k/8k/10k)	MLQA and TyDiQA-GoldP (1k/1k/5k)	Japanese Source (1k/1k/5k)
ATS	WikiLingua (131k/5k/5k)	WikiLingua (1k/1k/5k)	WikiLingua (1k/1k/5k)
DG	Our Previous work: Race++DG (135k/17k/17k)	We Created: HiDG (1k/1k/5k)	---

Dataset: Continued

- Auxiliary task: 11k/11k/11k (English/Hindi/Japanese)
- We Created high Quality Hindi DG (HiDG) dataset
- It took approximately 120 man-hours by two native Hindi speakers
- Data preparation Steps:
  a. Extracted <passage, question, answer> triplet from English SQuAD1.1 >150 tokens
  b. Generated distractors using our existing system (CIKM'20 work)
  c. Generated distractors were translated to Hindi using Google Translator
  d. Manually verified the quality by human annotators.

Baselines

- MT Pipeline (mBART): (a) M1= mBART + task-specific English data.
  (b) Te = GoogleTranslate(target language test data)
  (c) Te_out = M1(Te)
  (d) Traget_lang_out = GoogleTranslate(Te_out)

- mBART+MADMO: mBART + auxiliary task (Masking And Denoising objective with Mono-lingual data in three languages)

- mBART+MADPD: Inspired from (Chi et al., 2020), we took Parallel Data (English-Hindi and English-Japanese) and concatenate each parallel instances of two languages. Did same as previous.

Evaluation Metrics

Automatic Evaluation Metric

1. BLEU (sacrebleu)
2. ROUGE (1, 2 & L)
3. BERTScore

Manual Evaluation Metric

1. Fluency (1-5)
2. Relatedness (1-5)
3. Correctness (1-5)
4. Distractibility (1-5)

Automatic Evaluation Results: Hindi

Model	News Headline Generation			Question Generation			Abstractive TS			Distractor Generation		
Metrics	**R-1**	**R-2**	**R-L**	**BL**	**R-L**	**BS**	**R-1**	**R-2**	**R-L**	**BL**	**R-L**	**BS**
Cross-lingual zero-shot generation results												
MT Pipeline(mBART)	16.61	4.91	15.83	2.6	21.31	71.53	11.15	3.11	10.93	1.6	9.66	67.35
mBART+MADMO	29.32	16.36	27.52	3.9	23.70	73.76	18.25	4.92	16.10	2.8	15.86	72.26
mBART+MADPD	24.02	13.41	23.29	4.3	25.29	73.74	10.47	2.55	12.30	2.9	15.43	72.89
ZmBART	**34.94**	**19.38**	**32.74**	**4.4**	**26.51**	**74.19**	**21.27**	**5.30**	**17.64**	**4.1**	**21.05**	**73.39**
Cross-lingual few-shot generation results (with 1000 supervised data points)												
ZmBART	52.37	35.52	50.50	7.6	34.11	78.29	36.29	14.21	27.22	6.5	26.58	78.27

Table 1: Zero and few-shot cross-lingual generation results for Hindi Language

Automatic Evaluation Results: Japanese

Model	News Headline Generation			Question Generation			Abstractive TS		
Metrics	**R-1**	**R-2**	**R-L**	**BL**	**R-L**	**BS**	**R-1**	**R-2**	**R-L**
Cross-lingual zero-shot generation results									
MT Pipeline(mBART)	13.82	0.38	7.92	8.9	26.92	71.93	17.90	3.98	18.46
mBART+MADMo	33.75	8.12	17.78	16.6	34.80	74.01	28.74	9.01	23.63
mBART+MADPd	31.58	6.98	18.95	18.2	36.22	74.99	19.17	4.89	18.22
ZmBART	**35.25**	**9.24**	**19.92**	**18.8**	**38.74**	**75.91**	**36.60**	**15.26**	**29.85**
Cross-lingual few-shot generation results (with 1000 supervised data points)									
ZmBART	47.06	22.36	31.55	30.4	53.98	82.66	41.65	20.33	33.49

Table 2: Zero and few-shot cross-lingual generation results for Japanese Language

Manual Evaluation

Fluency: How fluent the generated text is?
Relatedness: How much generated outputs are in the context with input?
Correctness: It measures semantics and meaningfulness.
Distractibility: The degree of confusion for generated incorrect options.

- For both Hindi and Japanese 50 data point are taken from ATS, QG and NHG and 100 are taken for DG
- 60+ native Hindi/Japanese human annotator and 15 days evaluation duration

Manual Evaluation Results: Hindi

Model	News Headline Generation			Question Generation			Abstractive TS			Distractor Generation		
Metrics	Flu	Rel	Corr	Flu	Rel	Corr	Flu	Rel	Corr	Flu	Rel	Dist
Annotator set-01												
mBART+MADMO	3.86	4.34	3.94	2.66	3.38	3.52	3.56	3.58	3.22	3.61	4.08	2.89
mBART+MADPD	2.54	2.96	2.28	3.1	3.4	3.78	2.26	2.62	1.92	2.42	3.72	3.08
ZmBART	**4.14**	**4.22**	**4.04**	**3.24**	**3.44**	**3.9**	**4.02**	**4.12**	**3.54**	**4.12**	**4.19**	**3.83**
Annotator set-02												
mBART+MADMO	3.84	4.18	3.8	3.83	4.63	3.96	3.38	3.96	3.4	3.38	3.00	2.24
mBART+MADPD	2.96	3.02	2.7	**3.98**	4.70	3.98	2.96	3.16	2.84	2.97	3.11	**2.46**
ZmBART	**4.12**	**4.38**	**4.16**	3.95	**4.80**	**4.27**	**4.24**	**4.52**	**4.38**	**3.56**	**3.18**	2.36
Annotator set-03												
mBART+MADMO	3.56	3.74	**3.78**	2.68	3.76	3.32	2.9	3.34	2.9	3.96	3.74	3.12
mBART+MADPD	3.1	3.42	2.91	2.80	3.88	3.56	2.64	2.34	2.46	4.13	3.74	2.94
ZmBART	**3.70**	**3.84**	3.76	**2.86**	**4.04**	**3.76**	**4.06**	**3.56**	**3.56**	**4.44**	**4.12**	**3.12**

Table 3: Manual evaluation results of Zero-shot generated outputs for Hindi language

Manual Evaluation Results: Japanese

Model	News Headline Generation			Question Generation			Abstractive TS		
Metrics	Flu	Rel	Corr	Flu	Rel	Corr	Flu	Rel	Corr
Annotator set-01									
mBART+MADMO	2.66	2.98	2.50	1.98	**3.70**	**3.18**	3.04	3.55	3.44
mBART+MADPD	2.26	2.70	2.04	2.00	3.38	2.82	1.44	2.22	2.20
ZmBART	**3.60**	**4.02**	**3.50**	**2.12**	3.30	2.94	**4.24**	**3.90**	**3.90**
Annotator set-02									
mBART+MADMO	2.1	2.58	1.98	1.24	1.70	1.33	2.56	3.40	2.62
mBART+MADPD	1.58	1.78	1.46	**1.46**	1.72	1.78	1.00	1.00	1.00
ZmBART	**3.78**	**4.16**	**3.86**	1.26	**1.76**	**1.88**	**4.04**	**4.26**	**3.84**
Annotator set-03									
mBART+MADMO	2.24	2.72	2.24	**2.34**	2.46	2.39	2.82	3.18	3.52
mBART+MADPD	1.9	2.14	1.82	2.10	2.66	2.28	1.16	1.84	1.44
ZmBART	**2.88**	**3.22**	**2.92**	2.10	**2.70**	**2.46**	**3.32**	**3.52**	**3.04**

Table 4: Manual evaluation results of Zero-shot generated outputs for Japanese language

Results Analysis: Supervised Training Results

Task	Setting	BL	R-1	R-2	R-L	BS
NHG	W/ Aux-Task	15.9	43.22	21.33	40.88	90.13
	W/O Aux-Task	15.9	43.15	21.25	40.77	90.13
QG	W/ Aux-Task	20.6	53.20	26.53	51.37	92.18
	W/O Aux-Task	21.4	52.66	26.63	51.25	92.41
ATS	W/ Aux-Task	16.0	40.01	18.11	38.29	90.2
	W/O Aux-Task	15.8	39.52	18.00	37.91	90.10
QG	W/ Aux-Task	10.3	31.76	14.89	31.18	89.33
	W/O Aux-Task	10.0	31.87	14.59	31.30	89.42

Table 5: Automatic evaluation results of mBART on task-specific supervised English dataset (with and without Auxiliary Task)

Results Analysis: Does auxiliary task leads to spurious solution rather generalization ability?

- The generated headlines don't contain large continuous sequences from input text
- QG and DG are more challenging tasks and have objectives vastly different from the Auxiliary task's objective.
- Incorporation of auxiliary task improves the performance of diverse downstream tasks on real benchmark datasets, and does not favor any specific task or dataset.

Results Analysis: Effect of Catastrophic Forgetting

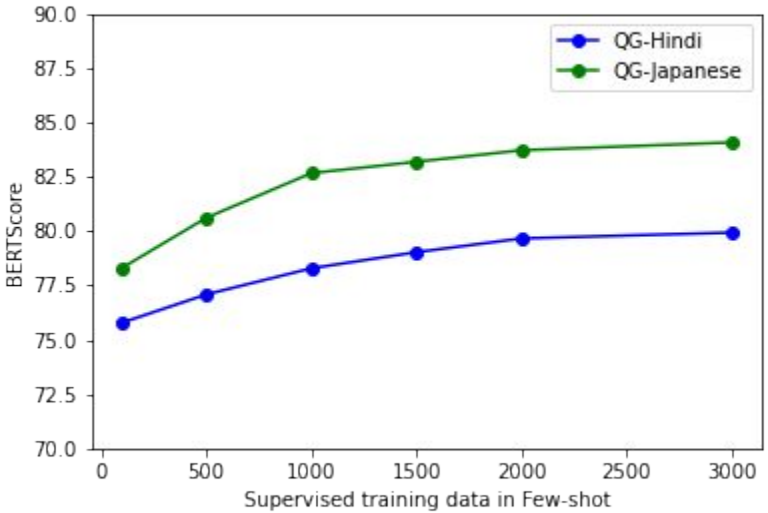
Setup	Setting-Details	BL(hi/ja)	R-L(hi/ja)	BS(hi/ja)
Model Components	Freeze word embedding (WE)	2.5/13.6	21.55/31.99	72.02/73.18
	Freeze WE + subset of Encoder & Decoder layers	2.9/15.3	22.62/36.60	72.24/72.98
	Freeze WE + Encoder layers	2.2/13.8	19.69/36.91	69.73/72.97
	Freeze WE + Decoder layers (ZmBART)	**4.4/18.8**	**26.51/38.74**	**74.19/75.91**
Regularized Optimization	Elastic Weight Consolidation (EWC)	2.1/11.6	18.21/29.47	68.36/72.91

Table 6: Evaluation score for different modeling approaches to deal catastrophic-forgetting for QG Task

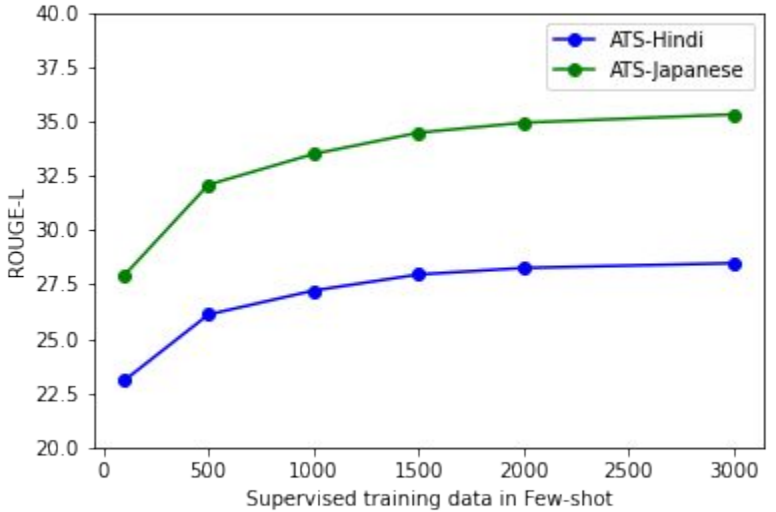
Setup	Setting-Details	R-1(hi/ja)	R-2(hi/ja)	R-L(hi/ja)
Model Components	Freeze word embedding (WE)	13.02/26.07	05.67/03.96	12.45/17.62
	Freeze WE + subset of Encoder & Decoder layers	14.27/25.72	06.70/03.21	13.76/18.28
	Freeze WE + Encoder layers	09.81/22.67	04.10/02.38	09.66/13.68
	Freeze WE + Decoder layers (ZmBART)	**34.94/35.25**	**19.38/09.24**	**32.74/19.92**
Regularized Optimization	Elastic Weight Consolidation (EWC)	12.01/22.16	05.43/03.11	11.22/16.31

Table 7: Evaluation score for different modeling approaches to deal catastrophic-forgetting for NHG Task

Results Analysis: Few-shot performance with Supervised data:



Question Generation



Abstractive text Summarization

Conclusion:

- In this paper, we propose a novel unsupervised framework (ZmBART) for cross-lingual transfer and generation
- The framework transfers supervision from HRL to LRLs which enables zero-shot language generation.
- The framework does not use any direct or pseudo-parallel data
- We performed experiments in three languages and 18 task-setup combinations: four supervised tasks in English, four tasks in Hindi (each with zero and and few shot), and three tasks in Japanese (each with zero and few shot).
- In future we want to extend this work by adding multiple other languages and tasks, and also explore other choices of auxiliary tasks for better model transfer.