CHARSPAN: Utilizing Lexical Similarity to Enable Zero-Shot Machine Translation for Extremely Low-resource Languages

Kaushal Kumar Maurya[1,3] and Rahul Kejriwal[2]
Maunendra Sankar Desarkar[1] and Anoop Kunchukuttan[2]
[1]NLIP Lab, IIT Hyderabad, India
[2]Microsoft, India [3]MBZUAI, UAE
Email: `cs18resch11003@iith.ac.in`

EACL 2024

Introduction

- Ethnologue list existence of over 7000 languages, but only around 300 languages has wikipedia articles.
- Most NLP research focuses on English only [1, 2] - less inclusive and less diverse.
- Many languages lack parallel or monolingual data and are not represented in existing multilingual PLMs/LLMs, termed Extremely Low Resource Languages or ELRLs.
- ELRs are resource-constrained subsets of low-resource languages (LRLs).

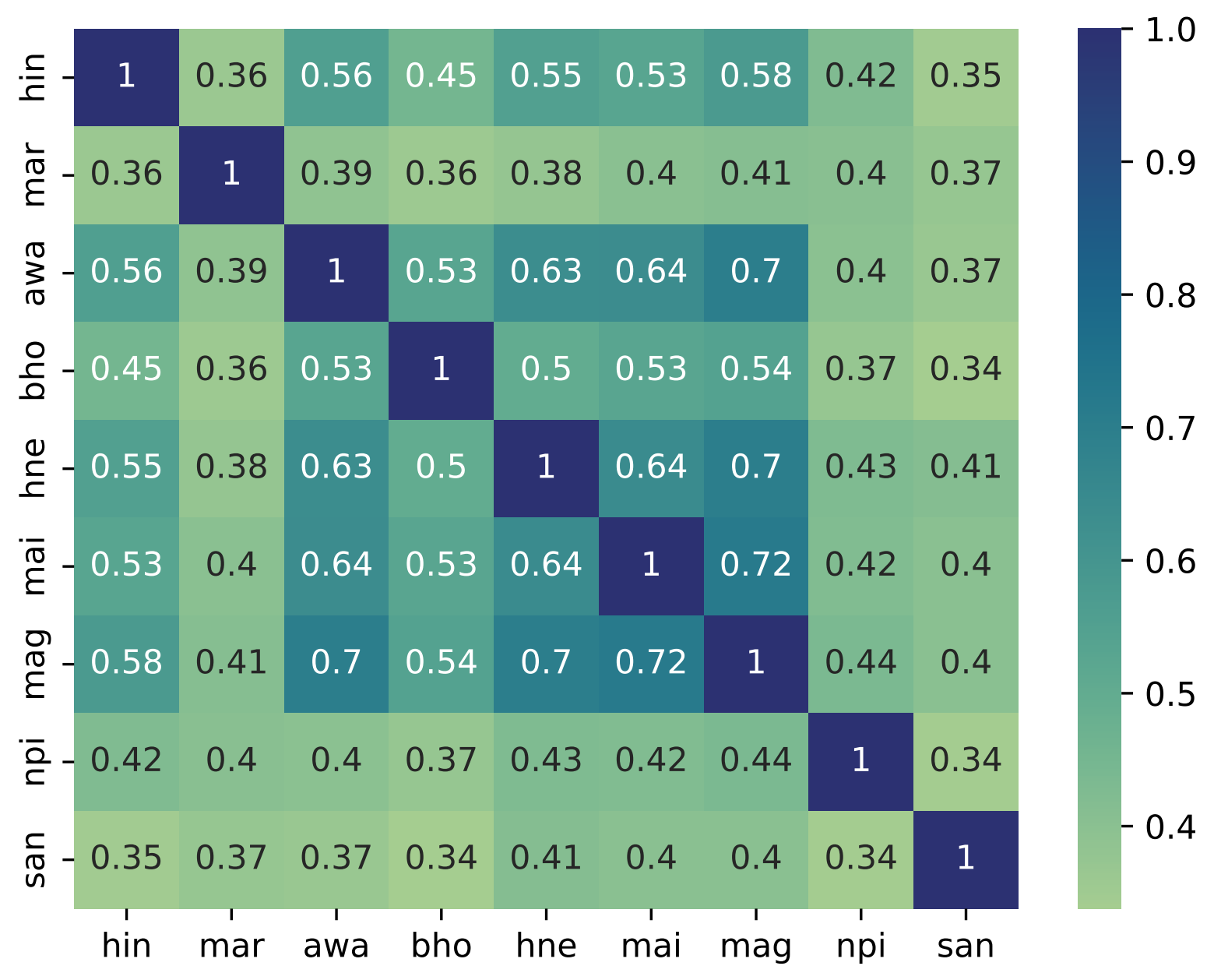
Motivation

Observation: Many ELRLs are lexically similar to some high-resource languages (HRLs) due to dialectal variations, vocabulary sharing, and geographical proximity. For example, Bhojpuri (an ELRL) is lexically very similar to Hindi (an 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 Hindi and Bhojpuri languages



Lexical similarity heatmap

Potential Modeling Direction:

- Utilize surface-level lexical similarity between HRLs and LRLs in the modeling.
- Noise augmentation is a plausible direction. Where noise is injected in HRL's training data which acts as augmented training data for ELRLs.
- The idea has been around; for example, random unigram noise augmentation (UNA) [3] was explored. This is limited to NLU tasks and suboptimal for NLG tasks.
- We hypothesize that existing methods do not work well for ELRLs which are lexically distant from HRLs.
- To overcome these limitations, we propose CHARSPAN, a character span-based noise augmentation model for machine translation (MT). The CHARSPAN model requires only HRLs' alphabet and is applicable for distant languages.

HRL (HIN): इस सीज़न में बीमारी के शुरुआती मामले जुलाई के आखिर में सामने आए थे।
ENG: The initial cases of the disease this season were reported in late July.

HRL (HIN)+CSN: ए_ सीज़न म बीमारी के __प_ मामले जुलाई के आखिर म सामने आए _।

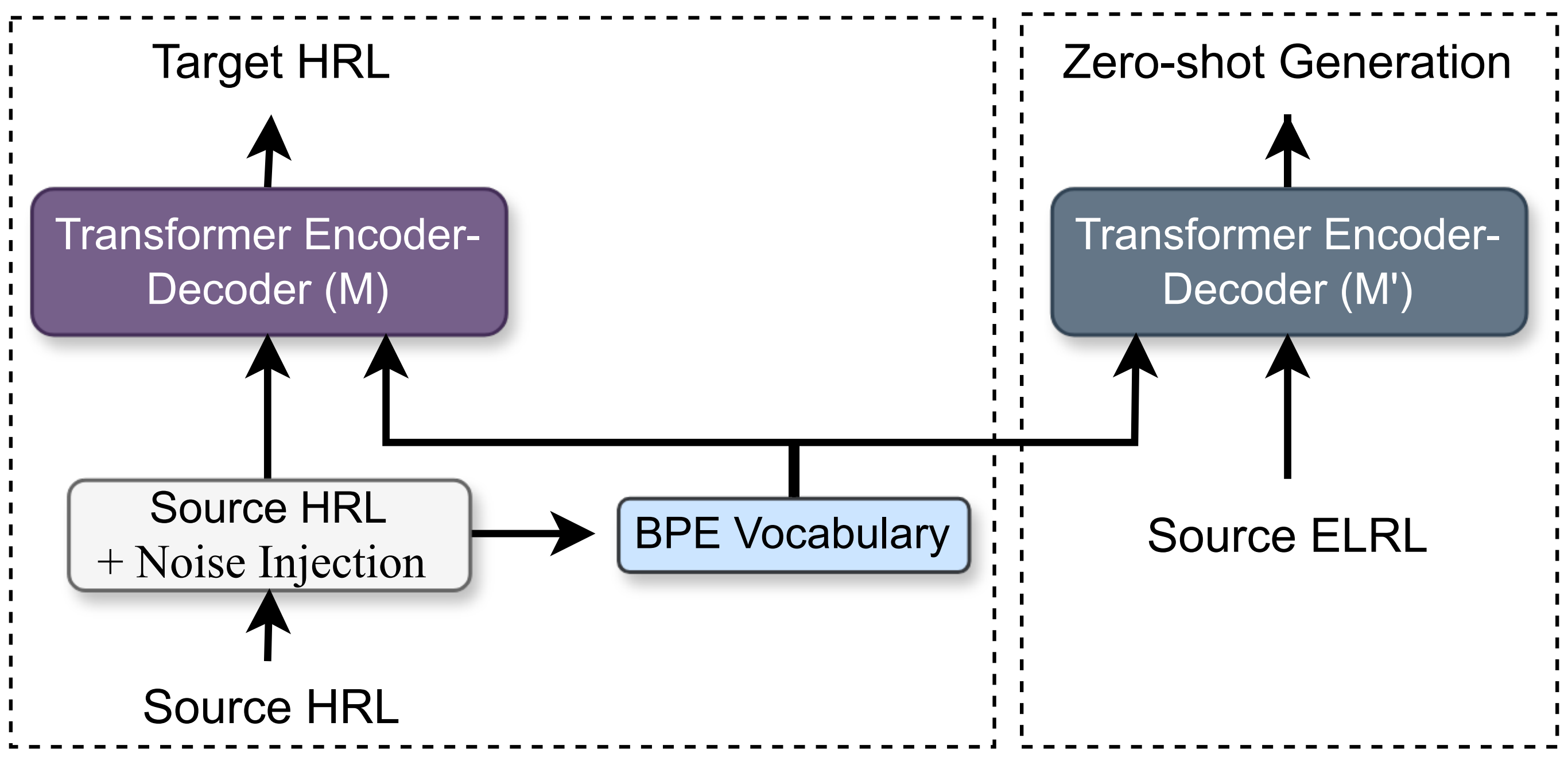
ELRL1 (BHO): ए सीजन में ई बीमारी क पहिला मामला जुलाई क आखिर में सामने आ गइल रहले।

ELRL2 (HNE): ए सीजन म ए बीमारी के पहिला मामला जुलाई के आखिर म सामने आए रहिस।

Problem Statement

Machine Translation (MT) from ELRLs → English in the *zero-shot* setting.

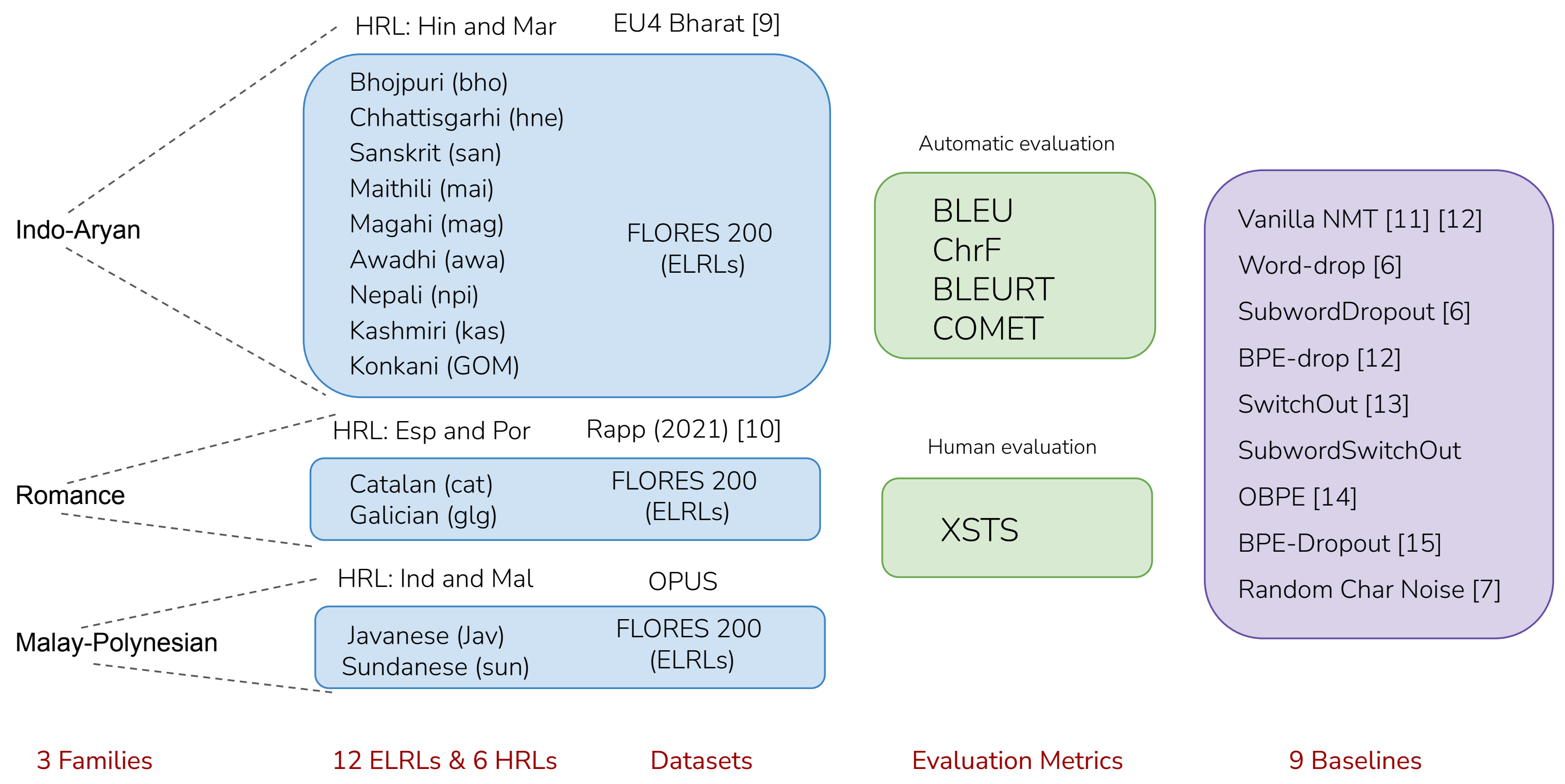
Proposed Methodology: CHARSPAN


Target HRL
Transformer Encoder-Decoder (M)
Source HRL + Noise Injection
BPE Vocabulary
Source HRL
Zero-shot Generation
Transformer Encoder-Decoder (M')
Source ELRL

(a) Supervised Training Phase (b) Generation Phase

- **Constraints:** HRLs and LRLs should be closely related.
- **Data Source:** No monolingual or parallel data for ELRLs. Used only HRL's alphabets.
- **Noise Augmentation:** The character span noise is augmented in the source side (HRL) of HRL to English parallel data. It acts as a augmented training data for ELRL → English MT task.
- **Selected Span:** We have performed random 1-3 character span noise augmentation.
- **Noise Augmentation Operations:** span deletion and span insertion (n-gram character span is replaced with a single character).
- **Model Training:** No pre-trained LLMs, trained from scratch.
- **Noise Injection Percentage:** randomly augment 10-11% characters for each input sequence.
- **Zero-shot Evaluation:** Trained on proxy data and evaluated with unseen ELRLs.
- **Intuition:** The noise injection acts as a regularizer, which accounts for lexical variations between HRL and LRLs. This improves the lexical similarity and cross-lingual transfer.

Experimental Setup



Results: ChrF Scores

Models	Indo-Aryan								Romance		Malay-Polynesian		Average
	Gom	Bho	Hne	San	Npi	Mai	Mag	Awa	Cat	Glg	Jav	Sun	
BPE*	26.75	39.75	46.57	27.97	30.84	39.79	48.08	46.28	33.32	53.75	31.44	32.21	38.06
WordDropout	27.01	39.57	46.19	28.13	31.91	40.31	47.37	46.48	34.20	52.21	32.03	32.52	38.16
SubwordDropout	27.91	40.11	46.26	29.46	32.56	40.99	47.91	47.43	35.09	52.28	33.38	33.47	38.90
WordSwitchOut	25.17	38.81	45.87	26.21	29.95	39.69	47.53	44.54	32.98	51.81	31.84	32.49	37.24
SubwordSwitchOut	26.08	38.84	45.84	28.19	30.81	40.19	47.28	45.93	33.26	53.71	31.24	32.06	37.78
OBPE	27.90	40.57	47.46	28.52	31.99	40.71	49.10	47.16	32.33	52.77	29.98	30.88	38.28
SDE	28.01	40.91	47.88	28.66	32.03	40.82	48.96	47.30	33.72	53.95	31.84	31.24	38.77
BPE-Dropout*	28.65	40.84	46.58	28.80	31.88	40.79	47.86	47.32	34.56	55.83	32.01	32.97	39.00
unigram char-noise**	28.85	42.53	49.35	29.80	34.61	42.67	50.97	49.43	43.16	54.81	35.42	36.69	41.52
BPE → SpanNoise*** (*ours*)	28.66	41.94	49.48	30.49	35.66	44.75	50.55	49.21	43.11	54.89	36.12	37.11	40.16
CHARSPAN (*ours*)	29.71	43.75	51.69	**31.40**	36.52	45.84	51.90	50.55	43.51	55.46	36.24	37.31	42.82
CHARSPAN + BPE-Dropout (*ours*)	**29.91**	**44.02**	**51.86**	30.88	**37.15**	**46.52**	**52.99**	**51.34**	**44.93**	**55.87**	**36.97**	**38.09**	**43.37**

CharSpan improvements over these baselines are statistically significant with *($p < 0.0001$), **($p < 0.001$), and *** ($p < 0.05$).

Analysis: Performance for Lexically Less Similar Languages

Languages	BPE	Unigram Noise	Char-Span Noise	Sim
Gujarati	34.36	36.17	38.09	0.42
Punjabi	29.18	33.34	36.50	0.40
Bengali	25.35	28.42	30.28	0.34
Telugu	23.30	24.05	24.12	0.27
Tamil	13.81	13.69	14.40	0.15

Zero-shot chrF scores; script conversion; **HRL:** Hindi and Marathi; **Sim:** lexical similarity.

Conclusions

- We propose a novel CHARSPAN model based on character span noise augmentation to enable/improve zero-shot ELRLs → English MT. We have achieved consistent improvement across different language families and datasets.
- In the future, we will extend this study to English → ELRLs MT, other NLG tasks, and languages.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020.

[3] Noëmi Aepli and Rico Sennrich. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of Association for Computational Linguistics 2022*, Dublin, Ireland, May 2022.

Acknowledgements

This research was conducted during the first author's internship at Microsoft India. We thank Microsoft for the support. Additionally, we extend our appreciation to the anonymous reviewers and meta-reviewers for their constructive feedback, which greatly contributed to the refinement of this work.