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 and Information Processing (NLIP) Lab
Indian Institute of Technology Hyderabad, India

cs23resch01004@iith.ac.in

Introduction

- There are 7000+ languages (Ethnologue), with only 300 languages having a presence in Wikipedia
- Majority of NLP research focuses on English [1, 2] only - less inclusive and less diverse
- Large number of languages *lack parallel data*, have *limited monolingual data*, *no representations* in existing multilingual PLMs called Extremely Low Resource Languages (ELRLs)

Motivation

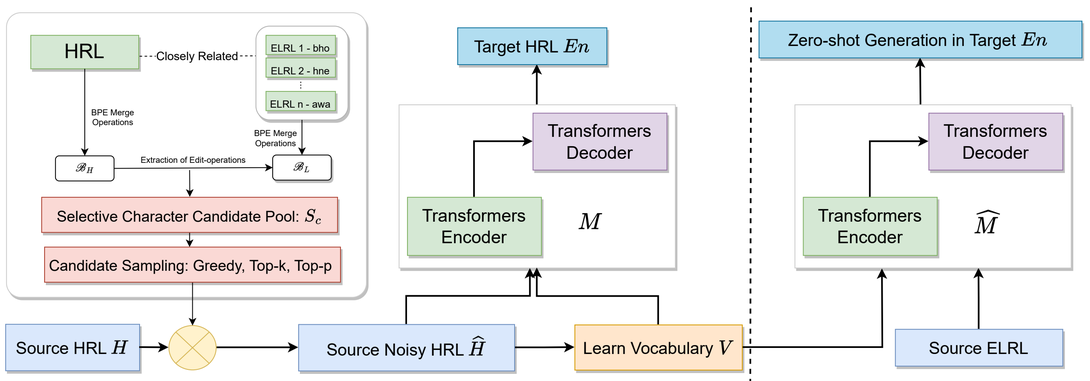
- **Hopeful Direction:** Many ELRLs are *lexically similar* to some HRLs due to dialectal variations, vocabulary sharing, and geographical proximity. For example, Bhojpuri (ELRL) is lexically very similar to Hindi (HRL).
- Utilize *surface-level lexical similarity* in generative modeling
- *Noise Injection* is a promising direction. For example, Random Noise Injection [3] explored for NLU task. However, it may be *suboptimal* for NLG tasks.
- Noising strategy should be systematic and incorporate *linguistic* signals.

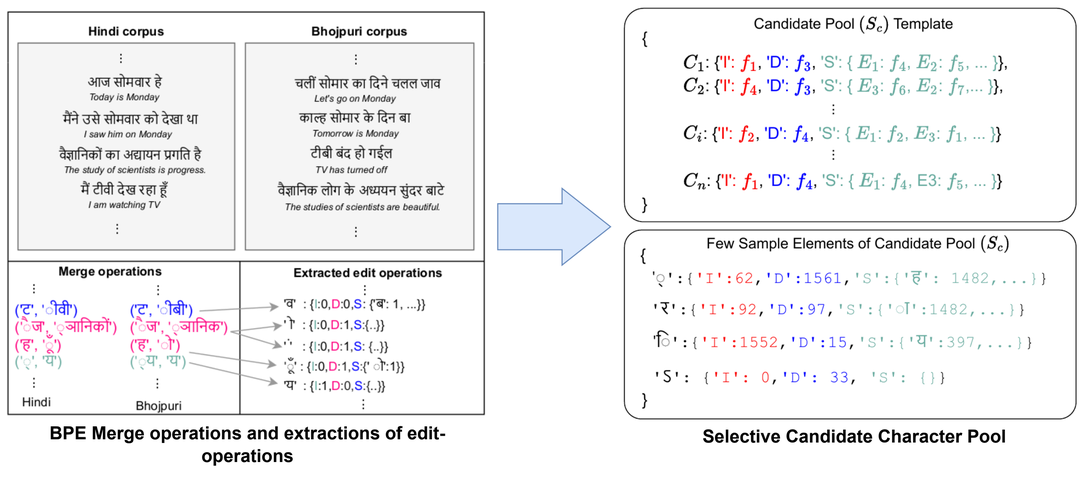
Problem Statement

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

Methodology: SELECTNOISE

- Selective Character Noise injection is performed in the source side (HRL) of HRL to English parallel data. It acts as *proxy* parallel training data for ELRL → English MT task.
- 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.
- We proposed a noise injection approach, SELECTNOISE, that is unsupervised, systematic and linguistically inspired.
- In SELECTNOISE, noise injection candidates are extracted with BPE merge operations and edit operations (called selective noise) in an unsupervised way.
- Noise is injected with sampling algorithm: greedy, top-k, and top-p.

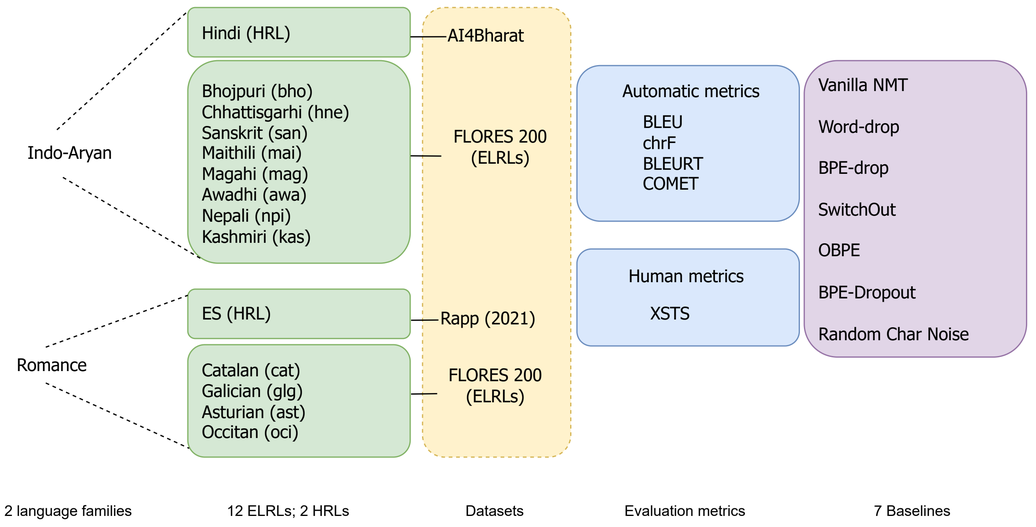




BPE Merge operations and extractions of edit-operations

Selective Candidate Character Pool

Experimental Setup



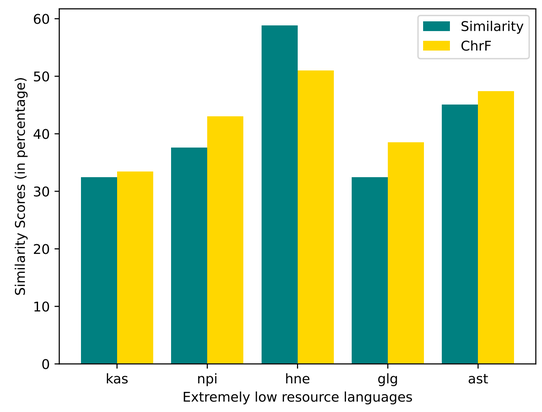
- Utilized ∼ 1000 examples from the monolingual corpus of each ELRL in SELECTNOISE
- Noise injection percentage of 5-10%
- Zero-shot setting: Training only on proxy HRL parallel data and evaluating with unseen ELRLs

Results

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

Analysis: Language Similarity vs. Performance



Conclusions

- We propose a novel SELECTNOISE that incorporates linguistics driven noise injection approach to improve zero-shot ELRLs → English.
- In the future, we will extend this study to more 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

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