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, Kaushal Kumar Maurya, and Maunendra Sankar Desarkar

Natural Language & Information Processing Lab (NLIP Lab)
Indian Institute of Technology Hyderabad, India

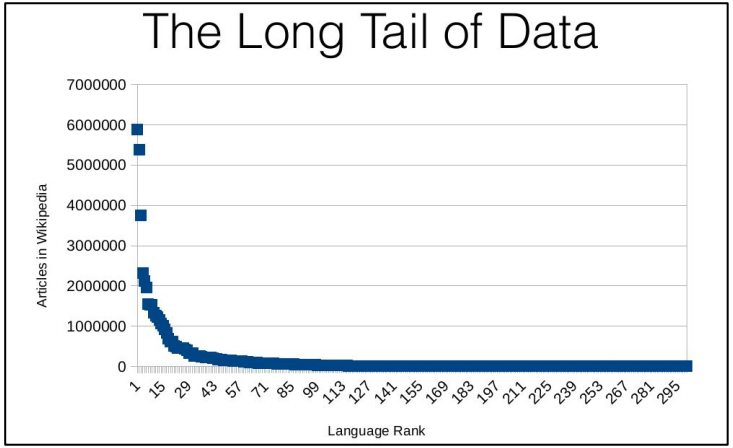
Slide link

Outline

- Introduction and Motivation
- Problem Statement
- Methodology
- Experimental Setup
- Results and Analyses
- Conclusion and Future Work

Introduction: Language Landscape

- 7000+ languages across the globe
- 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 [1, 2]



Source: Graham Neubig Multilingual NLP Lectures

Introduction: Machine Translation (MT)

- Cross lingual transfer among languages - Multilingual NMT [3]
- Reduce reliance of parallel data - Unsupervised NMT [4]
- Monolingual corpus incorporated NMT - Back-translation [5]
- Data augmentation approaches for MT:
  - word level perturbation [6]
  - overlapping BPE among related languages [8]

Introduction: ELRLs

Languages lack parallel data, have limited monolingual data, no existing multilingual pre-trained language models - Extremely Low Resource Languages (ELRLs)

Limited Efforts has been made for ELRL for MT task

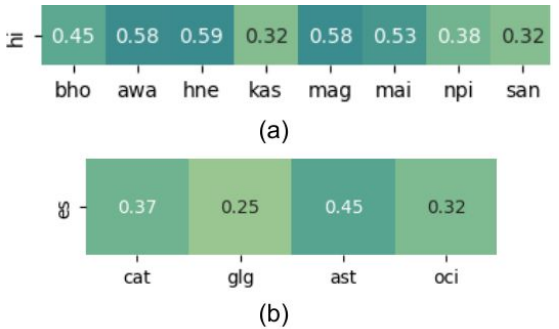
Motivation: Hopeful direction

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

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

hin:	**कनाडियन के खिलाफ नडाल का** सीधा **रिकॉर्ड 7-2** है।
bho:	**कनाडा के खिलाफ़ नाडाल के** हेड-टू -हेड **रिकॉर्ड 7-2** के बा।

Lexical level similarity between languages

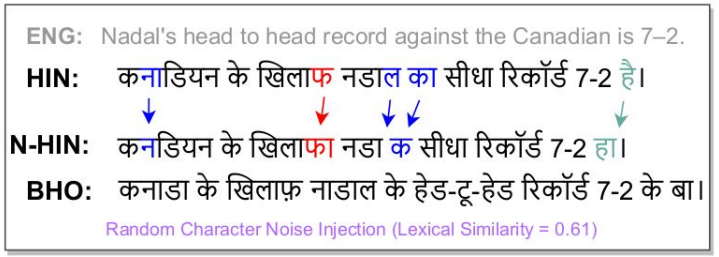


Lexical Similarity heatmap

Motivation: Hopeful direction

Earlier Successful for ELRL:

- Recall: Exploit lexical similarity through injecting random noise [2]
- Studies limited to NLU tasks only



Limitations:

- 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

Problem Statement

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

Proposed Methodology: Overview

Methodology:

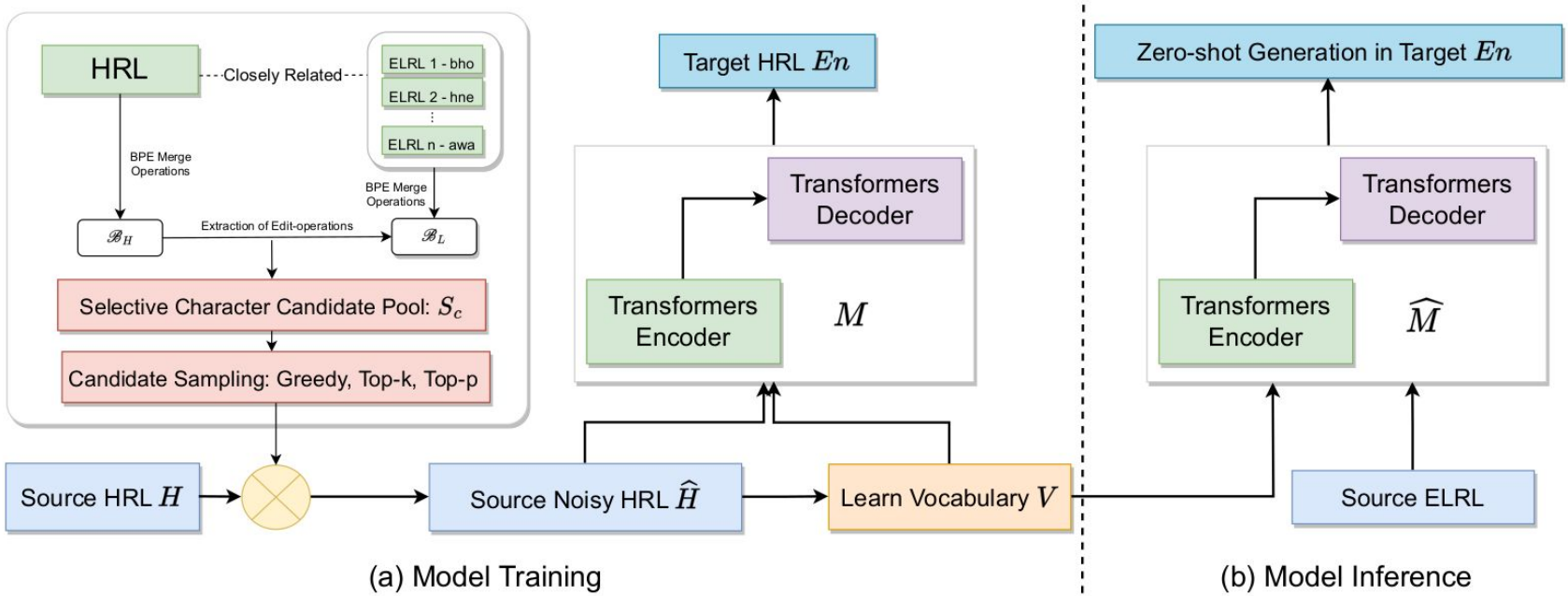
- Proposed character noise injection-based modeling approach
- Noise injection is performed in HRL to English parallel data
- Act as proxy parallel training data for ELRL to English translation task
- The noise injection candidates are extracted with BPE merge operations and edit operations (called selective noise)
- Noise is injected with sampling algorithm: Greedy, top-k and top-p

Proposed Methodology: Overview

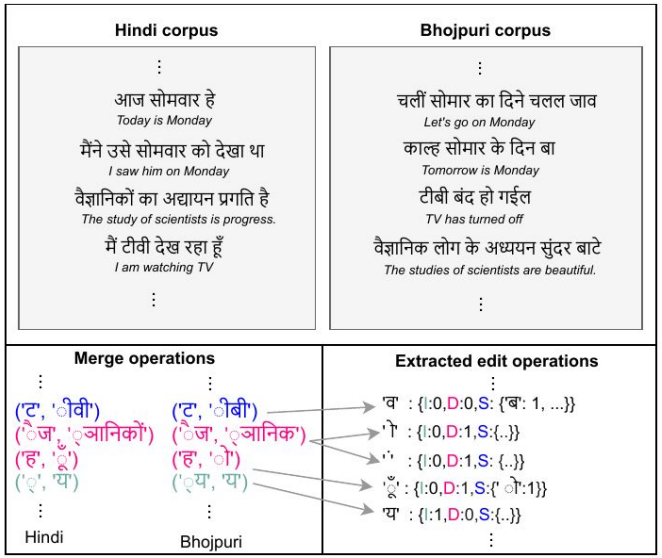
- **Intuition:**
  - Noise injection act as regularizer
  - Facilitate better a cross-lingual transfer from HRL to ELRL in source side

- **Hypothesis:**
  - Selective noise injection model is expected to outperform random noise injection
  - Performance of the selective noise injection should be comparable to supervised noise injection

Proposed Methodology: SELECTNOISE



Proposed Methodology: Candidate Extraction



BPE Merge operation and edit-operations

Candidate Pool (S_c) Template

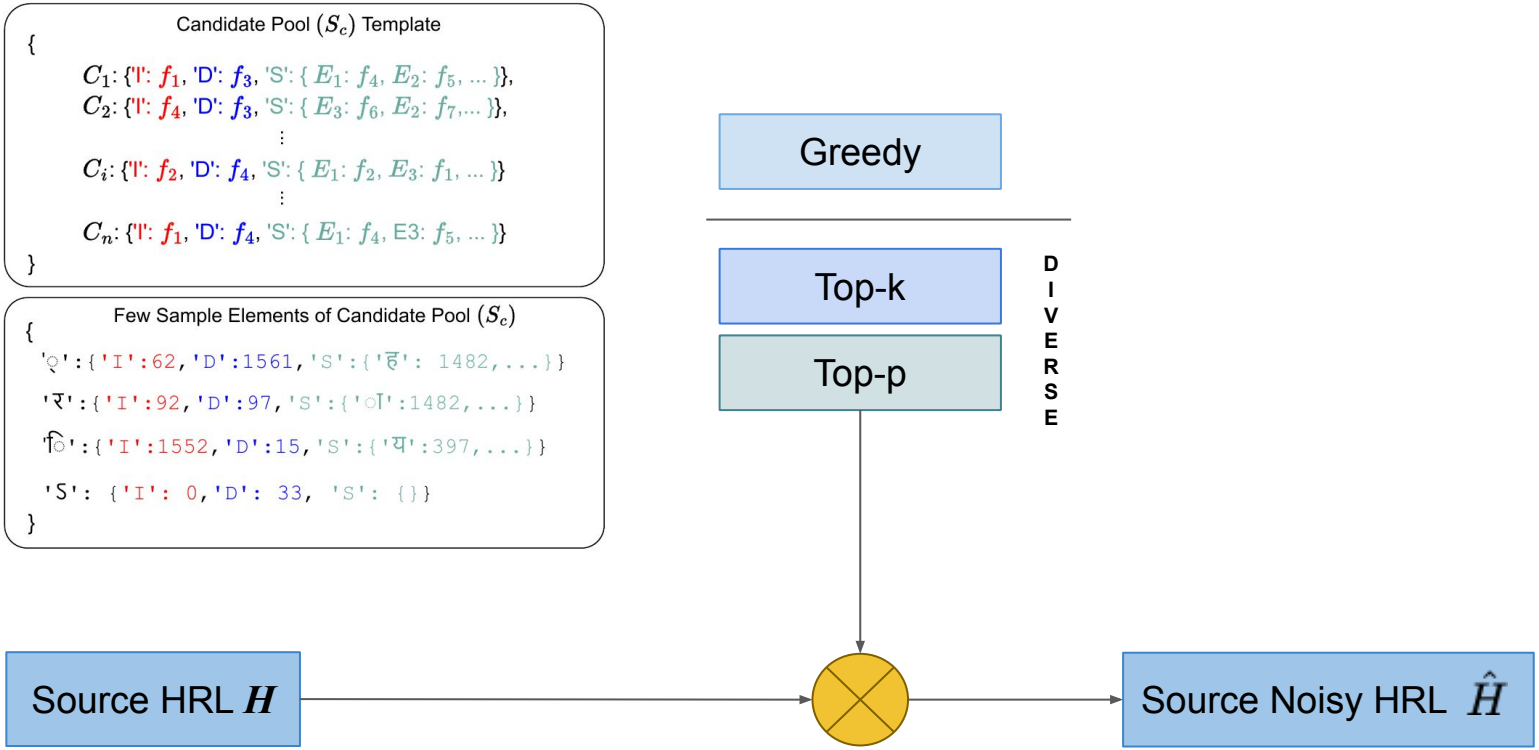
$$
\begin{aligned}
&\{\\
&\quad C_1\text{: \{'I': } f_1\text{, 'D': } f_3\text{, 'S': \{ } E_1\text{: } f_4\text{, } E_2\text{: } f_5\text{, ... \}\},}\\
&\quad C_2\text{: \{'I': } f_4\text{, 'D': } f_3\text{, 'S': \{ } E_3\text{: } f_6\text{, } E_2\text{: } f_7\text{,... \}\},}\\
&\quad \vdots\\
&\quad C_i\text{: \{'I': } f_2\text{, 'D': } f_4\text{, 'S': \{ } E_1\text{: } f_2\text{, } E_3\text{: } f_1\text{, ... \}\}}\\
&\quad \vdots\\
&\quad C_n\text{: \{'I': } f_1\text{, 'D': } f_4\text{, 'S': \{ } E_1\text{: } f_4\text{, E3: } f_5\text{, ... \}\}}\\
&\}
\end{aligned}
$$

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 Pooling

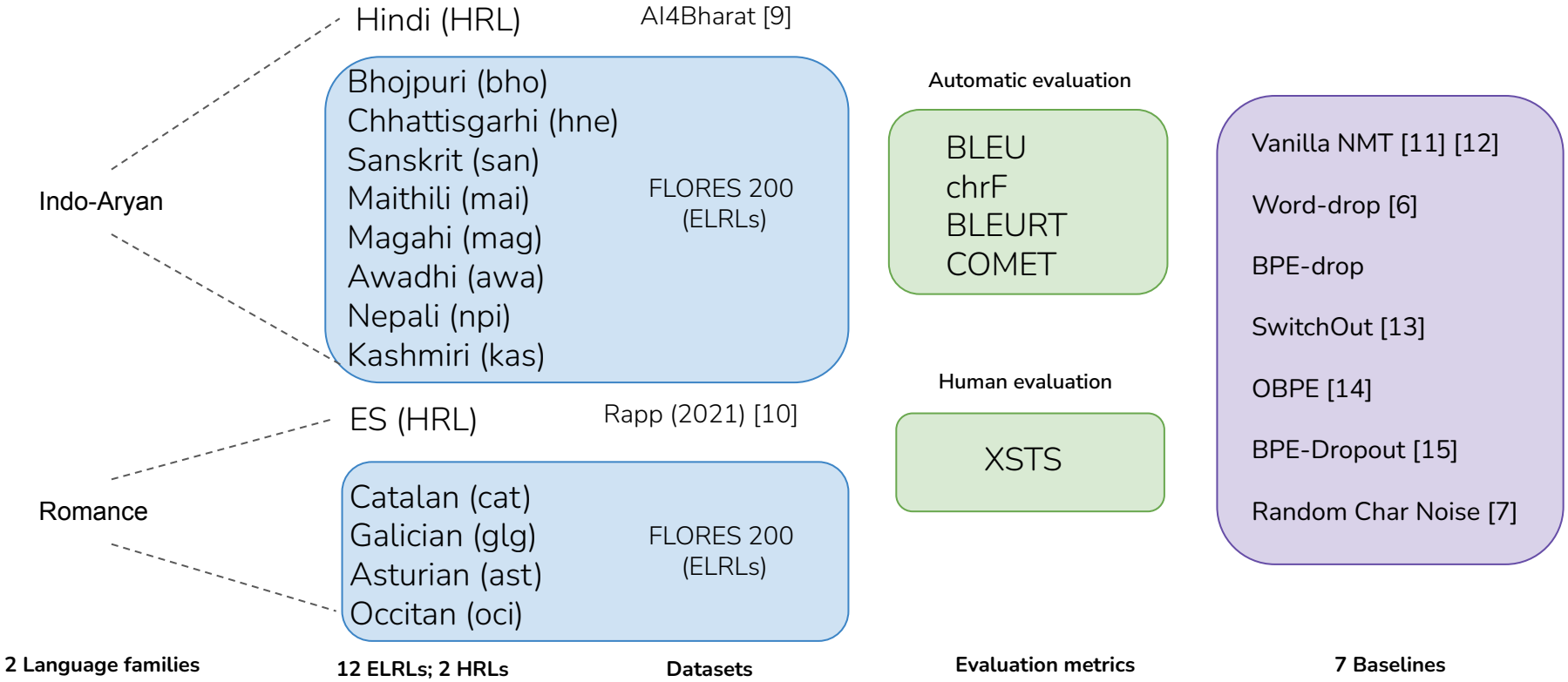
Proposed Methodology: Noise Injection



Experimental Setup

- ~ 1000 monolingual sentence are used for each ELRLs
- Noise injection percentage is 5-10% on related HRL data
- Zero-shot setting: Training only on proxy HRL parallel data and evaluate with unseen ELRLs

Experimental Setup



Results: Automatic Evaluation (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

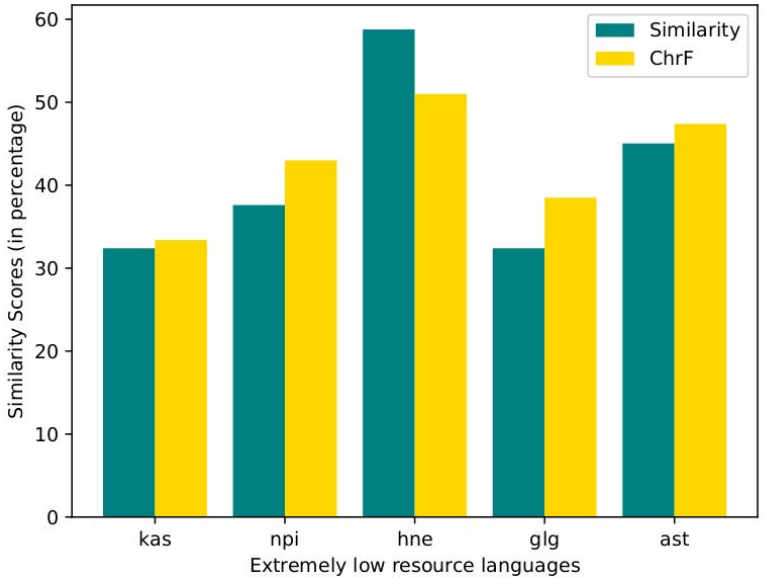
- Similar improvements in BLEU, COMET and BLEURT metrics

Results: Human evaluation

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) [16] metric scores between 1-5

Analysis: Language similarity vs Performance



Observation: High lexical similarities with High-resource language more the translation 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: Extracting edit-operations from larger monolingual corpus improves the translation performance

Conclusion & Future Work

- SelectNoise outperforms strong baselines across 12 ELRLs for ELRLs → English MT task
- Unsupervised noise injection gives comparable performance with Supervised approach
- Cumulative gain of 11.3% chrF over Vanilla-NMT

Future works:

- Extend to other NLG tasks
- Potential impact for English → ELRLs MT task

Acknowledgement

- Huge thanks to Human evaluators for assessment of translation performance

- Anonymous reviewers and Meta-reviewer for valuable insight and suggestions

- ACL Diversity & Inclusivity Grant

References

1. Emily Bender. 2019. The# benderrule: On naming the languages we study and why it matters. The Gradient, 14.
2. Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. 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, pages 6282–6293, Online. Association for Computational Linguistics.
3. Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In 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), pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics
4. Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In Proceedings of the Sixth International Conference on Learning Representations.
5. Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.
6. 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.
7. 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.
8. 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

9. Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK,Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. Transactions of the Association for Computational Linguistics, 10:145–162.
10. Reinhard Rapp. 2021. Similar language translation for Catalan, Portuguese and Spanish using Marian NMT. In Proceedings of the Sixth Conference on Machine Translation, pages 292–298, Online.
11. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.
12. 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.
13. 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.
14. 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.
15. 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.
16. Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

NATURAL LANGUAGE & INFORMATION PROCESSING LAB

EMNLP 2023

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

Thank you!!

Visit our lab page

Personal webpage

Contact us:

Mail: cs23resch01004@iith.ac.in
Lab Mail: nlip@cse.iith.ac.in

Lab Webpage: https://nlip-lab.github.io/