NATURAL LANGUAGE
& INFORMATION
PROCESSING LAB

EACL 2024

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

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

*Work conducted during first author's internship at Microsoft, India

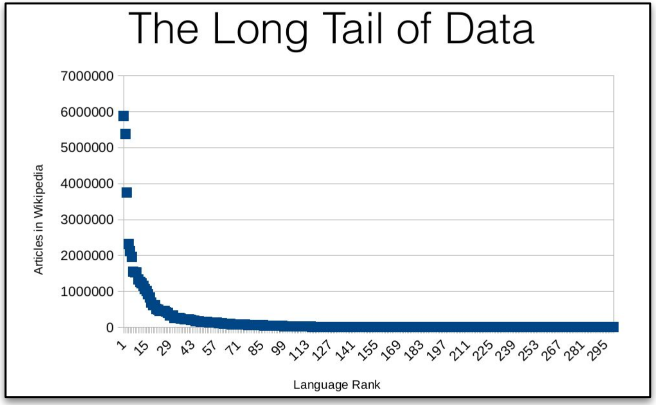
Download Slide

Outline

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

Introduction: Landscape of Low-resource Languages

- 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]



The Long Tail of Data

Most Spoken Languages of the World

1. English (1.132 B)
2. 中文(普通话) (1.116 B)
3. हिन्दी (615.4 M)
4. Español (534.4 M)
5. Français (279.8 M)
6. اَلْعَرَبِيَّةُ (273.9 M)
7. বাংলা (265.0 M)
8. Россия (258.2 M)
9. Português (234.1 M)
10. Bahasa Indonesia (279.8 M)

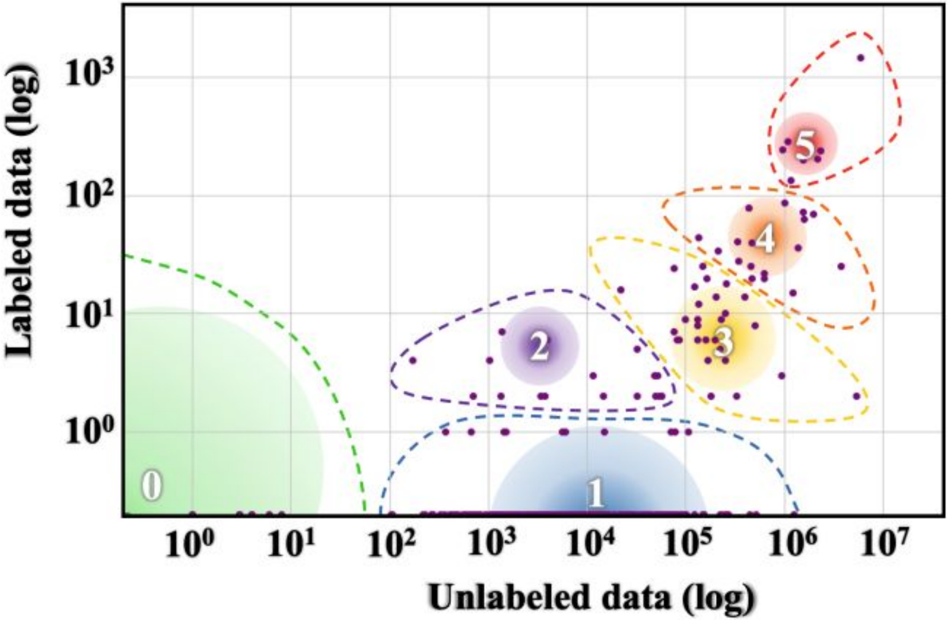
[3] Joshi et al., ACL 2020; [4] E. Bender, The Gradient, 2019

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

Introduction: Limited data for LRLs

- 88% languages fall into class 0 and untouched by language technology [3]

- Only ~100 languages are part of existing large language model, even for those languages, NLG (MT) adaptability is challenging [5]



[3] Joshi et al., ACL 2020; [5] Ahuja et al. 2023

Introduction: Extremely LRLs (ELRLs)

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

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:
    - ○ Word level perturbation
    - ○ BPE vocabulary overlapping among related languages [23]

Limited Efforts has been made for ELRL for MT task

[23] Patil et al., ACL 2022

Motivation: Hopeful direction

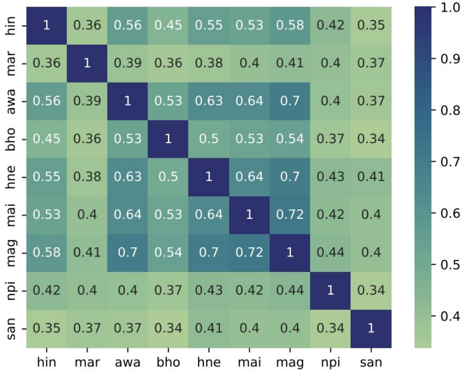
- Utilize relatedness among languages
  - Dialectal variations
  - Vocabulary sharing
  - Similarities due to Geographical proximity

- 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



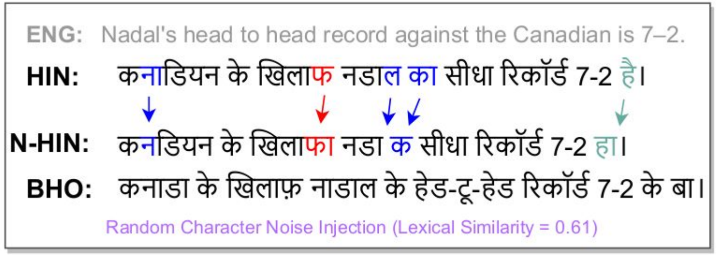
	hin	mar	awa	bho	hne	mai	mag	npi	san
hin	1	0.36	0.56	0.45	0.55	0.53	0.58	0.42	0.35
mar	0.36	1	0.39	0.36	0.38	0.4	0.41	0.4	0.37
awa	0.56	0.39	1	0.53	0.63	0.64	0.7	0.4	0.37
bho	0.45	0.36	0.53	1	0.5	0.53	0.54	0.37	0.34
hne	0.55	0.38	0.63	0.5	1	0.64	0.7	0.43	0.41
mai	0.53	0.4	0.64	0.53	0.64	1	0.72	0.42	0.4
mag	0.58	0.41	0.7	0.54	0.7	0.72	1	0.44	0.4
npi	0.42	0.4	0.4	0.37	0.43	0.42	0.44	1	0.34
san	0.35	0.37	0.37	0.34	0.41	0.4	0.4	0.34	1

Lexical Similarity heatmap

Motivation: Hopeful direction

Earlier Success for ELRL:

- Recall: Exploit lexical similarity through char-noise augmentation [24]



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।

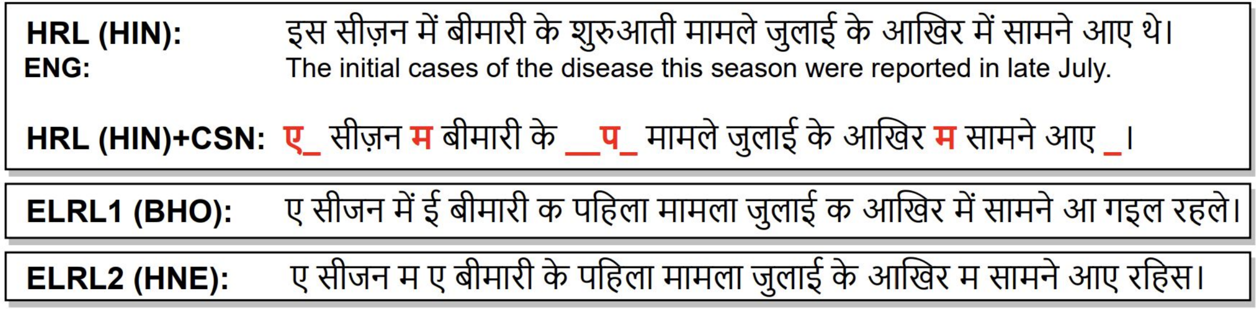
Random Character Noise Injection (Lexical Similarity = 0.61)

Limitations:

- Studies limited to NLU tasks only
- Applied with LLM vocab which hinders scalability
- Char Noise augmentation may be suboptimal

[24] Aepli et al., ACL 2022 (Findings)

Motivation: Beyond Character Noise Augmentation



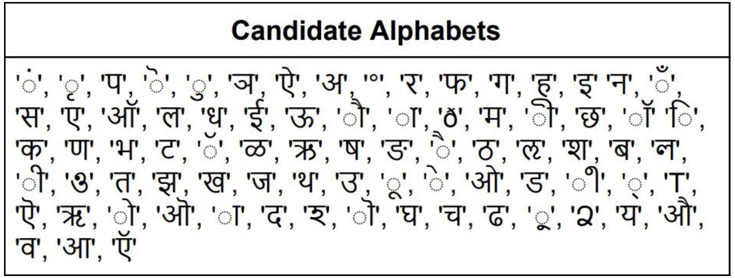
HRL (HIN): इस सीज़न में बीमारी के शुरुआती मामले जुलाई के आखिर में सामने आए थे।
ENG: The initial cases of the disease this season were reported in late July.

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।

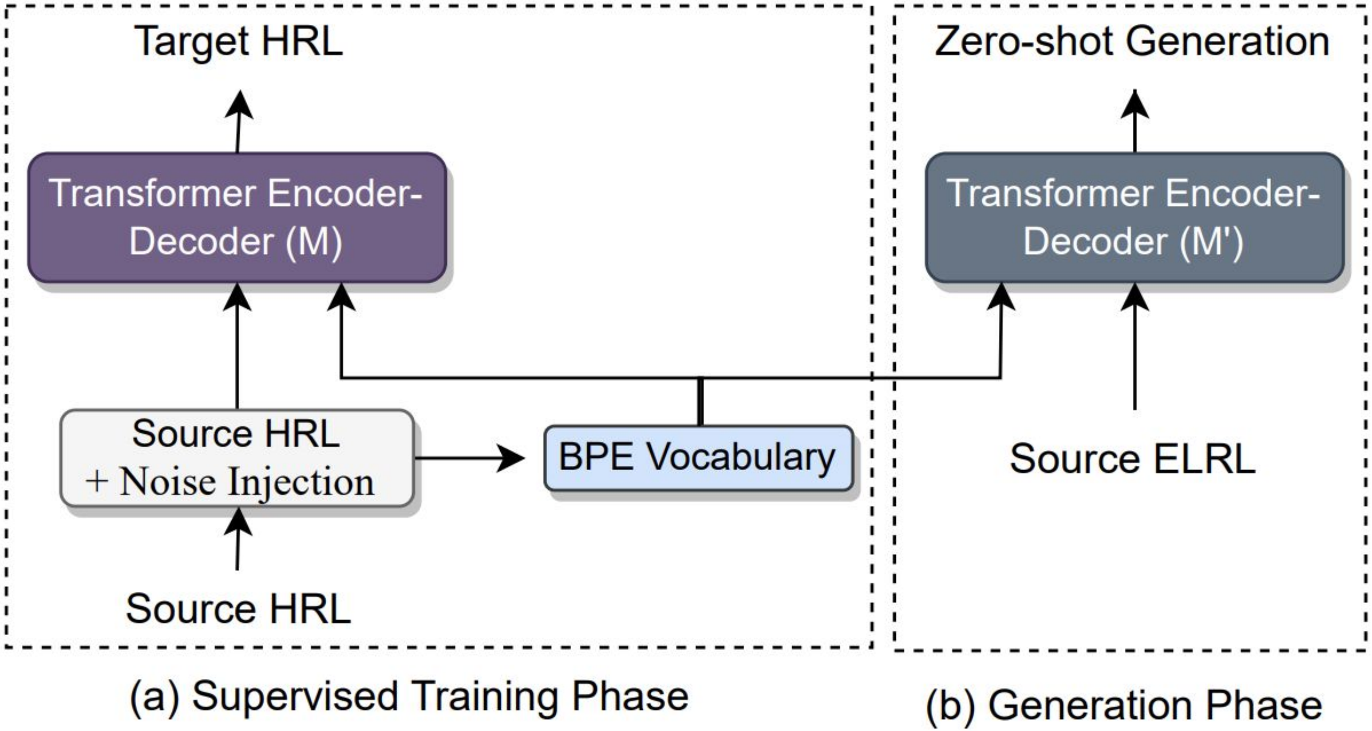
Character-span Noise Augmentation



Candidate Alphabets
'ं', 'ृ', 'प', 'े', 'ु', 'ञ', 'ऐ', 'अ', '॰', 'र', 'फ', 'ग', 'ह', 'इ' 'न', 'ँ', 'स', 'ए', 'ऑ', 'ल', 'ध', 'ई', 'ऊ', 'ौ', 'ा', 'ठ', 'म', 'ी', 'छ', 'ॉ' 'ि', 'क', 'ण', 'भ', 'ट', 'ॅ', 'ळ', 'ऋ', 'ष', 'ङ', 'ै', 'ठ', 'ऌ', 'श', 'ब', 'ल', 'ी', 'ও', 'त', 'झ', 'ख', 'ज', 'थ', 'उ', 'ू', 'ॆ', 'ओ', 'ड', 'ी', '्', 'T', 'ऍ', 'ॠ', 'ो', 'ऒ', 'ा', 'द', 'श', 'ॊ', 'घ', 'च', 'ढ', 'ू', '२', 'य', 'औ', 'व', 'आ', 'ऍ'

[24] Aepli et al., ACL 2022 (Findings)

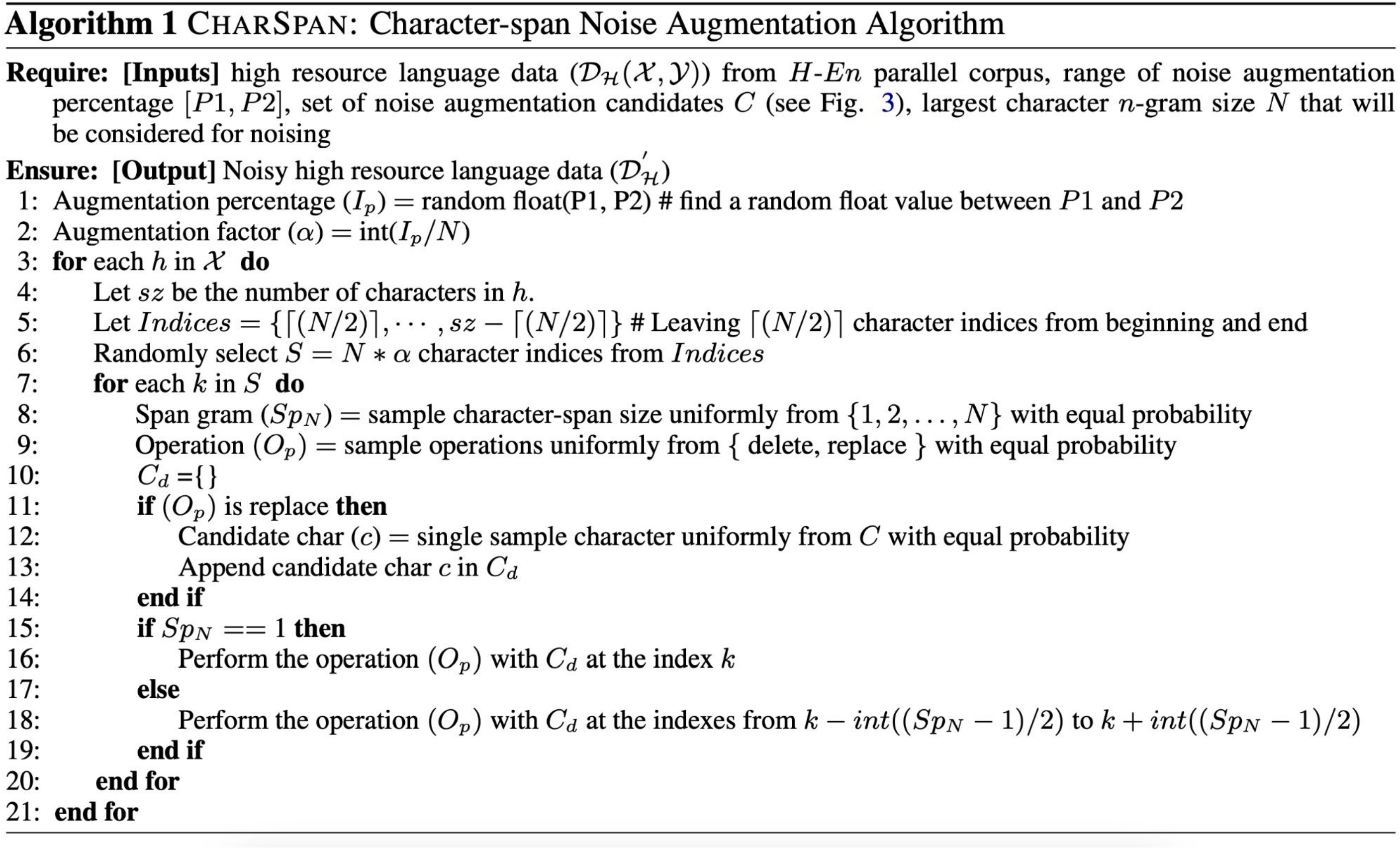
Methodology: CHARSPAN Model


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

Methodology: CharSpan Model

- Constraints: HRLs and LRLs should be closely related
- Data Sources:
  - No monolingual or parallel data for ELRLs.
  - Used only HRL's alphabets.
- Model Training: No pre-trained LLMs, trained from scratch.
- Noise Augmentation Span: Applied 1-3 character grams.
- Operations: Delete and n-gram to single character insertion.
- Noise Injection Percentage: Injected noise at 10-11%.
- Zero-shot Evaluation:
  - Trained on proxy HRL parallel data.
  - Evaluated with unseen ELRLs

Methodology: Algorithm



Algorithm 1 CHARSPAN: Character-span Noise Augmentation Algorithm

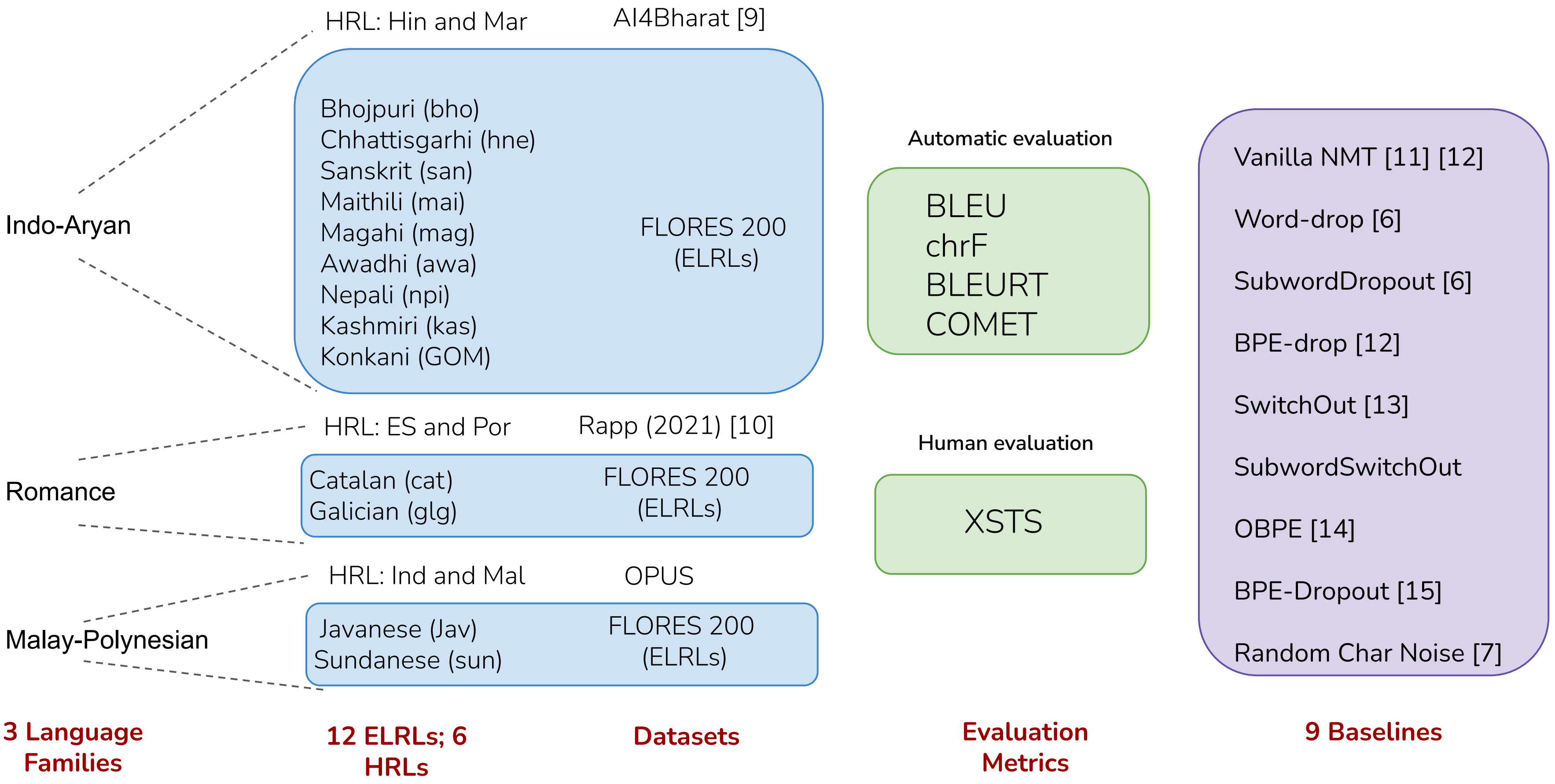
```
Require: [Inputs] high resource language data ($\mathcal{D}_{\mathcal{H}}(\mathcal{X}, \mathcal{Y})$) from $H$-$En$ parallel corpus, range of noise augmentation percentage $[P1, P2]$, set of noise augmentation candidates $C$ (see Fig. 3), largest character $n$-gram size $N$ that will be considered for noising
Ensure: [Output] Noisy high resource language data ($\mathcal{D}_{\mathcal{H}}^{'}$)
Augmentation percentage ($I_p$) $=$ random float(P1, P2) # find a random float value between $P1$ and $P2$
Augmentation factor ($\alpha$) $=$ int($I_p/N$)
for each $h$ in $\mathcal{X}$ do
    Let $sz$ be the number of characters in $h$.
    Let $Indices = \{\lceil(N/2)\rceil, \cdots, sz - \lceil(N/2)\rceil\}$ # Leaving $\lceil(N/2)\rceil$ character indices from beginning and end
    Randomly select $S = N * \alpha$ character indices from $Indices$
    for each $k$ in $S$ do
        Span gram ($Sp_N$) $=$ sample character-span size uniformly from $\{1, 2, \ldots, N\}$ with equal probability
        Operation ($O_p$) $=$ sample operations uniformly from { delete, replace } with equal probability
        $C_d$ ={}
        if ($O_p$) is replace then
            Candidate char ($c$) $=$ single sample character uniformly from $C$ with equal probability
            Append candidate char $c$ in $C_d$
        end if
        if $Sp_N == 1$ then
            Perform the operation ($O_p$) with $C_d$ at the index $k$
        else
            Perform the operation ($O_p$) with $C_d$ at the indexes from $k - int((Sp_N - 1)/2)$ to $k + int((Sp_N - 1)/2)$
        end if
    end for
end for
```

Methodology: Intuition

Intuition:

- Noise augmentation act as regularizer
- Facilitate better a cross-lingual transfer from HRL to ELRL in source side
- Char-Span Noise augmentation enable cross-lingual transfer to distant languages i.e., transfer to less lexically similar to HRLs

Experimental Setup



Indo-Aryan

HRL: Hin and Mar

AI4Bharat [9]

Bhojpuri (bho)
Chhattisgarhi (hne)
Sanskrit (san)
Maithili (mai)
Magahi (mag)
Awadhi (awa)
Nepali (npi)
Kashmiri (kas)
Konkani (GOM)

FLORES 200 (ELRLs)

Romance

HRL: ES and Por

Rapp (2021) [10]

Catalan (cat)
Galician (glg)

FLORES 200 (ELRLs)

Malay-Polynesian

HRL: Ind and Mal

OPUS

Javanese (Jav)
Sundanese (sun)

FLORES 200 (ELRLs)

Automatic evaluation

BLEU
chrF
BLEURT
COMET

Human evaluation

XSTS

Vanilla NMT [11] [12]

Word-drop [6]

SubwordDropout [6]

BPE-drop [12]

SwitchOut [13]

SubwordSwitchOut

OBPE [14]

BPE-Dropout [15]

Random Char Noise [7]

3 Language Families

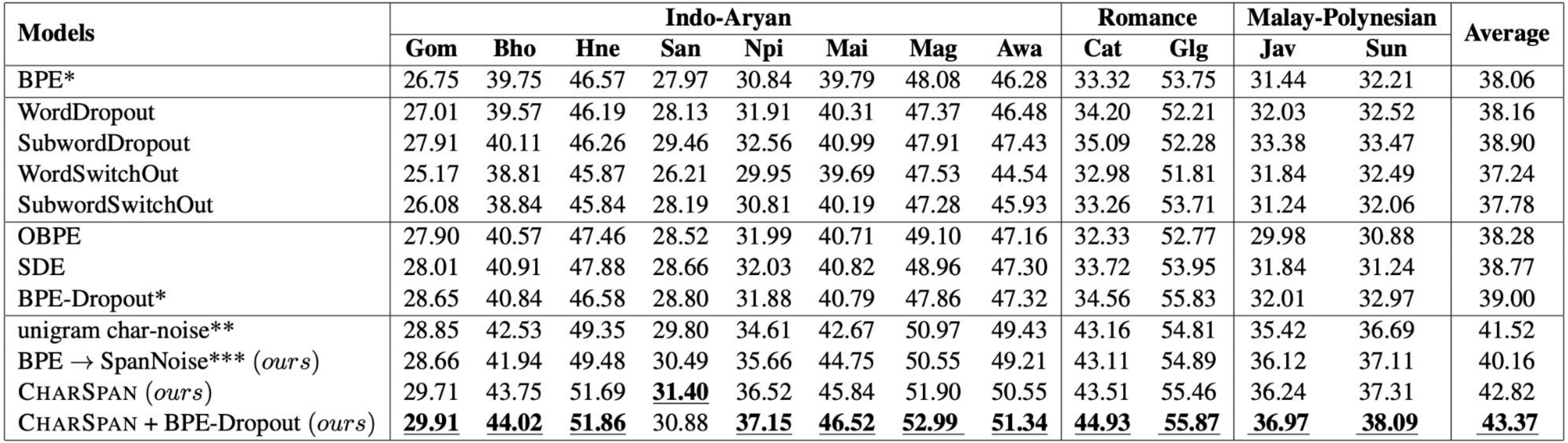
12 ELRLs; 6 HRLs

Datasets

Evaluation Metrics

9 Baselines

Evaluation Results [ChrF Scores]

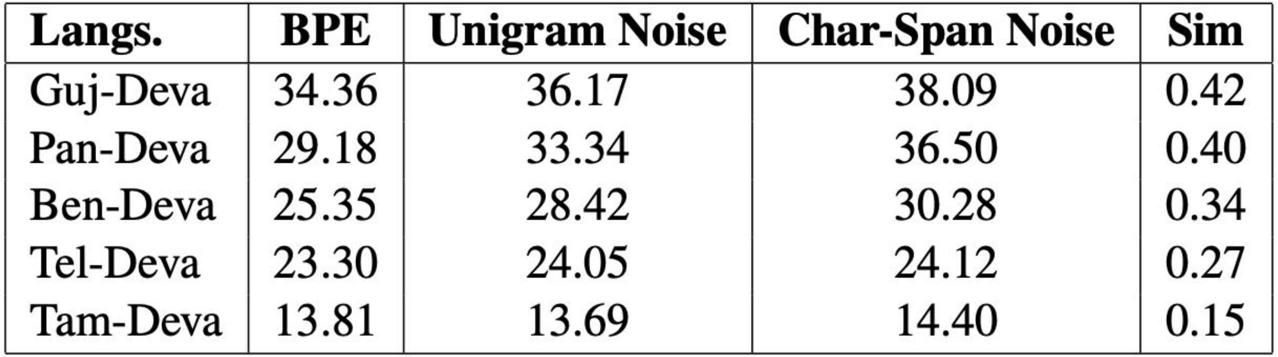


Models	Indo-Aryan								Romance		Malay-Polynesian		Average
	Gom	Bho	Hne	San	Npi	Mai	Mag	Awa	Cat	Glg	Jav	Sun	
BPE*	26.75	39.75	46.57	27.97	30.84	39.79	48.08	46.28	33.32	53.75	31.44	32.21	38.06
WordDropout	27.01	39.57	46.19	28.13	31.91	40.31	47.37	46.48	34.20	52.21	32.03	32.52	38.16
SubwordDropout	27.91	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**

Zero-shot chrF scores for ELRLs → English

- Similar improvements in BLEU, COMET and BLEURT metrics

Analysis: Performance for Distant Languages

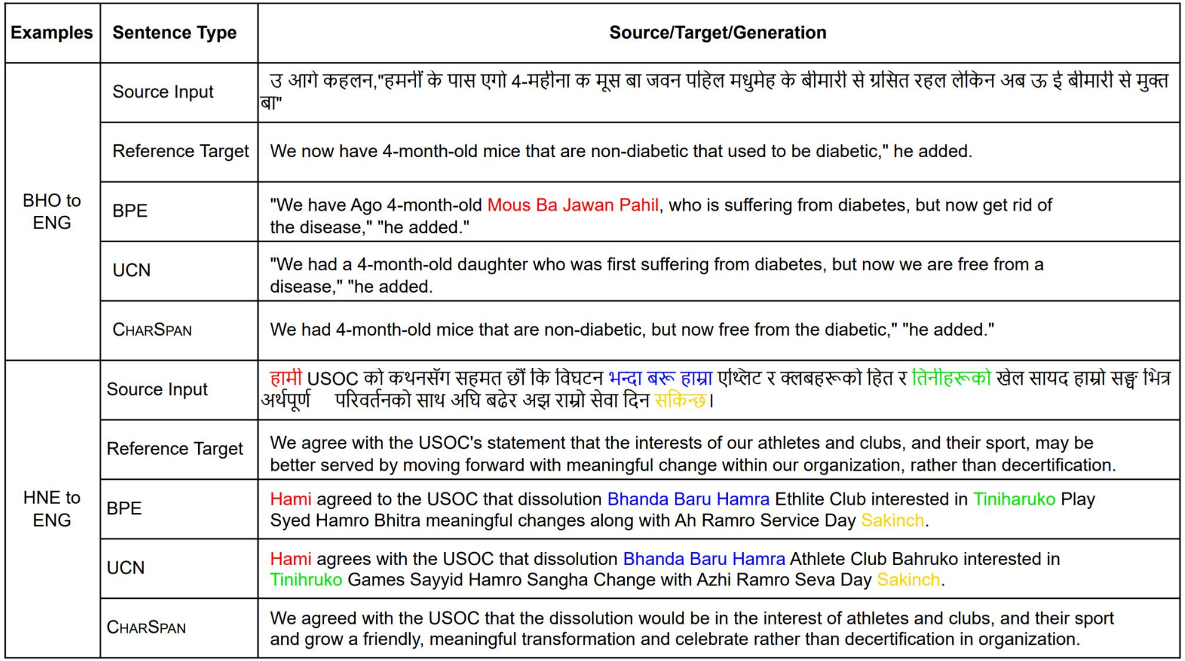


Langs.	BPE	Unigram Noise	Char-Span Noise	Sim
Guj-Deva	34.36	36.17	38.09	0.42
Pan-Deva	29.18	33.34	36.50	0.40
Ben-Deva	25.35	28.42	30.28	0.34
Tel-Deva	23.30	24.05	24.12	0.27
Tam-Deva	13.81	13.69	14.40	0.15

HRL are Hindi and Marathi. Sim: LCS similarity on char level

Observation: The Char-Span model has responsible performance even for distant languages.

Analysis: Mitigate Zero-shot Translation Errors



Examples	Sentence Type	Source/Target/Generation
BHO to ENG	Source Input	उ आगे कहलन,"हमनीं के पास एगो 4-महीना क मूस बा जवन पहिल मधुमेह के बीमारी से ग्रसित रहल लेकिन अब ऊ इ बीमारी से मुक्त बा"
	Reference Target	We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.
	BPE	"We have Ago 4-month-old Mous Ba Jawan Pahil, who is suffering from diabetes, but now get rid of the disease," "he added."
	UCN	"We had a 4-month-old daughter who was first suffering from diabetes, but now we are free from a disease," "he added.
	CHARSPAN	We had 4-month-old mice that are non-diabetic, but now free from the diabetic," "he added."
HNE to ENG	Source Input	हामी USOC को कथनसँग सहमत छौं कि विघटन भन्दा बरू हाम्रा एथ्लेट र क्लबहरूको हित र तिनीहरूको खेल सायद हाम्रो सङ्घ भित्र अर्थपूर्ण परिवर्तनको साथ अघि बढेर अझ राम्रो सेवा दिन सकिन्छ।
	Reference Target	We agree with the USOC's statement that the interests of our athletes and clubs, and their sport, may be better served by moving forward with meaningful change within our organization, rather than decertification.
	BPE	Hami agreed to the USOC that dissolution Bhanda Baru Hamra Ethlite Club interested in Tiniharuko Play Syed Hamro Bhitra meaningful changes along with Ah Ramro Service Day Sakinch.
	UCN	Hami agrees with the USOC that dissolution Bhanda Baru Hamra Athlete Club Bahruko interested in Tinihruko Games Sayyid Hamro Sangha Change with Azhi Ramro Seva Day Sakinch.
	CHARSPAN	We agreed with the USOC that the dissolution would be in the interest of athletes and clubs, and their sport and grow a friendly, meaningful transformation and celebrate rather than decertification in organization.

Observation: Char-Span Model Successfully mitigate the translation error from BPE and UNC models.

Conclusion & Future Work

- CharSpan Model outperforms strong baselines across 12 ELRLs for ELRLs → English MT task
- The proposed model does not required monolingual data, parallel data and LLM multilingual representation.
- Highly Scalable
- Cumulative gain of 12.34% chrF over Vanilla-NMT (BPE) model

Future works:

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

Acknowledgement

- Special thanks to Microsoft India for the internship opportunity and mentorship support.
- Gratitude to the anonymous reviewers and meta-reviewer for valuable insights and suggestions.

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: cs18resch11003@iith.ac.in
Lab Mail: nlip@cse.iith.ac.in

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