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

Anonymous ACL submission

Abstract

We address the task of machine translation (MT) from extremely low-resource language (ELRL) to English by leveraging cross-lingual transfer from *closely-related* high-resource language (HRL). The development of an MT system for ELRL is challenging because these languages typically lack parallel corpora and monolingual corpora, and their representations are absent from large multilingual language models. Many ELRLs share lexical similarities with some HRLs, which presents a novel modeling opportunity. However, existing subword-based neural MT models do not explicitly harness this lexical similarity, as they only implicitly align HRL and ELRL latent embedding space. To overcome this limitation, we propose a novel, CHARSPAN, approach based on *character-span noise augmentation* into the training data of HRL. This serves as a regularization technique, making the model more robust to *lexical divergences* between the HRL and ELRL, thus facilitating effective cross-lingual transfer. Our method significantly outperformed strong baselines in zero-shot settings on closely related HRL and ELRL pairs from three diverse language families, emerging as the state-of-the-art model for ELRLs.

1 Introduction

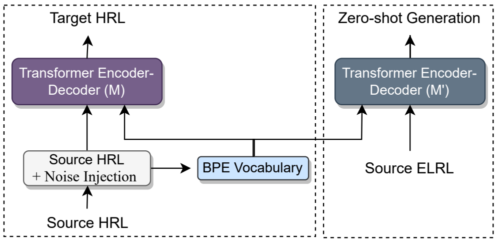
Recent advancements in multilingual modeling have expanded the coverage of Natural Language Processing (NLP) technologies to many LRLs by transferring knowledge from HRLs to LRLs. As a result, this progress has led to remarkable advancement in multiple NLP tasks, including MT, transliteration, natural language understanding, and text generation (Johnson et al., 2017; Kunchukuttan et al., 2018; Conneau et al., 2020; Liu et al., 2020) for LRLs. However, most of the existing work has focused on the top few hundred languages represented on the web (Joshi et al., 2020b). The availability of monolingual corpora and/or parallel corpora for these languages has been the driving force behind this progress, achieved either through direct training, few-shot training, or learning with large multilingual language models (mLLMs). This enables learning common embedding spaces that facilitate cross-lingual transfer (Nguyen and Chiang, 2017; Khemchandani et al., 2021). However, there is a long tail of languages for which no monolingual or parallel corpora are available, and they are absent from mLLMs. These languages are referred to as ELRLs. This paper is a step toward building MT systems for ELRLs.

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।

Figure 1: Hindi (HIN; HRL), Bhojpuri (BHO; ELRL) and Chhattisgarhi (HNE; ELRL) parallel sentences. Additionally, the corresponding noisy Hindi example with character-span (CSN) noise. BHO and HNE are closely related to HIN.

Fortunately, many of ELRLs are lexically similar to some HRLs. *Lexical similarity refers to languages sharing words with similar form (spelling and pronunciation) and meaning.*[1] This includes cognates, lateral borrowings and loan words. We explore if cross-lingual transfer can be enabled or improved for ELRLs by *explicitly* taking lexical similarity into account. In particular, *we explore MT from an ELRL to another language (English) with transfer enabled by a related HRL on the source side.* Our key *insight* is that cognates in ELRL having similar spelling to the HRL word can be thought of as misspellings of the latter. For example, the word लगता (*lagta*) in Hindi (HRL) is spelled as लागअता (*laagata*) in Bhojpuri (LRL). If we make the HRL model robust to spelling variations, it will improve cross-lingual transfer to related ELRLs. To achieve spelling variation robustness, we propose novel *character-span noise augmentation (CSN)* in the HRLs training data. A sample example is presented in Fig. 1. This

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


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

Figure 2: Overview of proposed CharSpan model

acts as a regularizer and makes the model more robust to perturbations in representations of words in closely related languages and improves model generalization for lexically similar languages.

Our key contributions are: (1) We propose a novel model CharSpan: Character-Span *noise augmentation*, which considers surface level lexical similarity to improve cross-lingual transfer between closely-related HRLs and LRLs. The proposed approach shows a 12.5% chrF improvement over baseline NMT models across all considered ELRLs. Our model also shows performance improvement over various data augmentation baselines. (2) We show that our approach generalizes across three typologically diverse language families, comprising 6 HRLs and 12 ELRLs. (3) We provide detailed ablation and analysis to gain insights and demonstrate the effectiveness of our approach.

2 Related Work

Traditionally, character-level noise has been used to improve the robustness of MT systems to spelling mistakes and ASR errors (Sperber et al., 2017; Vaibhav et al., 2019; Karpukhin et al., 2019). However, these approaches are mostly investigated for their impact on robustness rather than for cross-lingual transfer. More recently, token/BPE-level general noise augmentation approaches such as WordDropout (Sennrich et al., 2016a) and SwitchOut (Wang et al., 2018) have been proposed, but they have limited cross-lingual transfer capabilities. Close to our work, Aepli and Sennrich (2022) and Blaschke et al. (2023) show that augmenting data with character-level noise can help cross-lingual transfer. The models were evaluated with NLU tasks. n contrast, our work focuses on MT, an NLG task, which is much more challenging than an NLU task in a zero-shot setting. Furthermore, we explore span noise augmentation, which considers larger lexical divergence (less lexical similarity between the HRL and ELRL) and enables better cross-lingual transfer.

On utilizing lexical similarity literature thread, Patil et al. (2022) proposed OverlapBPE, which takes lexical overlap between HRL and LRL while learning BPE vocabulary. Provilkov et al. (2020) introduced BPE-Dropout, providing on-the-fly non-deterministic segmentations while training. Soft Decoupled Encoding (SDE) Wang et al. (2019) utilizes lexical information without pre-segmenting the data by decoupling the lexical and semantic representations. These models require small monolingual data for modeling. In contrast, the CharSpan model does not require any training resources for ELRLs other than language alphabets.

3 The CharSpan Model

Figure 2 presents an overview of the proposed CharSpan model, for ELRL to English MT task. The model has two phases: supervised training with noisy HRL and zero-shot generation with ELRLs.

Model Training and Generation: In the *supervised training phase*, the source-side training data of the HRL pair ($\mathcal{D}_{\mathcal{H}}$) is augmented with character-span noise (described later) to create the augmented parallel corpus ($\mathcal{D}^{'}_{\mathcal{H}} = \eta(\mathcal{D}_{\mathcal{H}})$), where η is the noise function. $\eta(\mathcal{D}_{\mathcal{H}})$ can be considered as the proxy parallel data for the ELRL-English translation task. Next, we learn a subword vocabulary ($\mathcal{V}$) using $\mathcal{D}^{'}_{\mathcal{H}}$, i.e., the noise is augmented before learning the vocabulary. A standard encoder-decoder transformer model ($\mathcal{M}$; Vaswani et al. (2017)) is then trained with $\mathcal{D}^{'}_{\mathcal{H}}$ and $\mathcal{V}$ from scratch in a supervised setting to obtain the trained model $\mathcal{M}^{'}$. Finally, in the *zero-shot generation phase*, for a given source ELR language $\mathcal{L}$, the target English translation is obtained using $\mathcal{M}^{'}$ and $\mathcal{V}$ in the zero-shot setting.

Character Span Noise (CSN) Function: The noise functions serve to make the model robust to spelling variations between related languages. This acts as a regularizer and helps improve cross-lingual representation and transfer. Intuitively, the existing unigram character noise might address limited lexical variations between HRL and ELRLs. *To address larger lexical divergence, we propose a CSN where span noise is augmented*. Formally, for a given sentence, $x \in \mathcal{X}$ from $\mathcal{D}^{'}_{\mathcal{H}}(\mathcal{X}, \mathcal{Y})$ with indices $I = 1, 2, \ldots, |x|$, a subset of these indices $I_s \subset I$ is randomly and uniformly selected as the starting point for the noise augmentation. Subsequently, 1-3 character gram spans are iteratively sampled until the noise augmentation budget (i.e., 9% - 11% characters) is exhausted. We employ *span deletion* and *span replacement with a single random character of ELRL*, both with equal probability as the noising operations[2]. This CSN is inspired by SpanBERT (Joshi et al., 2020a)[3]. A formal algorithm is presented in the Appendix. We conducted experiments with all three operations (including insertion), with different percentages of noise and various other experimental setups, as outlined in Appendix Table 13. We found the presented noise augmentation configuration to be the most effective.

4 Experimental Setup

We seek answers to the following questions: (1) Does the augmentation of CSN improve cross-lingual transfer, i.e., zero-shot performance for related ELRLs for MT task? (2) Why does the model's cross-lingual transfer improve? - Insights from the learned embedding space. (3) Is the proposed approach scalable to typologically diverse language families?

4.1 Datasets and Languages

We evaluated the performance of the proposed model on three language families: Indo-Aryan, Romance, and Malay-Polynesian. We considered six HRLs and twelve LRLs (two HRLs and several ELRLs from each family). All the ELRLs are lexically similar and have the same script with corresponding HRLs, as shown in Figure 4 (Appendix I). Parallel training data for the HRLs was selected from publicly available datasets. The model's performance was evaluated on the FLORES-200 devtest set (Costa-jussà et al., 2022). Dataset statistics are presented in Appendix C.

4.2 Baselines and Evaluation Metrics

Based on recent literature in low-resource MT, we compare our approach with the following strong baselines: (a) Vanilla NMT with BPE segmentation (BPE; Sennrich et al. (2016b)), (b) General data augmentation methods: (Sub)WordDropout and (Sub)WordSwitchOut, (c) Methods using lexical similarity: Overlap BPE, BPE-Dropout, SDE and unigram char-noising (Aepli and Sennrich, 2022). Baselines and model training details are provided in Appendix. Following recent studies on MT for ELRLs (Costa-jussà et al., 2022; Siddhant et al., 2022), we use chrF (Popović, 2015) as the primary evaluation metric. In addition, we also report BLEU (Papineni et al., 2002) and two neural metrics viz., BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020) scores in Appendix E.

5 Results and Analyses

The proposed CHARSPAN and baseline models' results across different language families are presented in Table 1. The following are the major observations:

Noise vs. Baselines: All the proposed noise augmentation models outperform vanilla NMT and all baseline models that utilize lexical similarity (i.e., OBPE, BPE-Dropout, and SDE). This trend is consistent across all language families and ELRLs. Moreover, existing lexical similarity-based baselines do not provide any major improvement in translation quality over vanilla NMT. Possible reasons for this can be twofold: (1) most of the ELRLs either do not have monolingual data (OBPE and SDE are required) or have small data, and (2) we observe that in OBPE, approximately 90% of vocabulary tokens are already overlapping among HRLs and ELRLs, leaving little room for learning additional overlapping tokens. This is expected, as these two language sets are closely related. The proposed CHARSPAN method also outperforms general data augmentation methods like (Sub)WordDropout and (Sub)WordSwitchout, showing its effectiveness.

Unigram vs. Char-Span Noise: We are first to explore unigram char noise (Aepli and Sennrich, 2022) for related language MT. We see that unigram char noise is beneficial for the task. However, our proposed CHARSPAN provides significant improvements over unigram character noise. We believe our proposed data augmentation is more effective in bringing language representations closer.

When to introduce noise? To understand when noise augmentation is effective, we augmented noise after learning the vocabulary in the baseline (BPE → CSN). This leads to improved performance over all baselines. This enables scalability since augmenting noise after learning the vocabulary allows the application of this method to large language models which have fixed vocabulary. However, the results suggest that applying noise prior to learning the vocabulary,

[2] We explored some linguistically motivated noising schemes, but these were not beneficial.

[3] SpanBERT applies denoising to subword tokens while we apply it at the character level.

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 → CSN (*our*)	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 (*our*)	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 (*our*)	**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**

Table 1: Zero-shot chrF scores results for ELRLs → English machine translation.

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

Table 2: Zero-shot chrF scores with additional lexically less similar languages. `HRL`: hi and mr; `sim`: lexical similarity

Models	Bho	Hne	San	Npi	Mai	Mag	Awa
BPE	0.761	0.793	0.701	0.744	0.762	0.809	0.792
UCN	0.853	0.888	0.765	0.821	0.849	0.897	0.883
CharSpan	**0.871**	**0.909**	**0.789**	**0.858**	**0.868**	**0.913**	**0.901**

Table 3: Average cosine similarity between representations of source HRLs and source ELRLs for Indo-Aryan family. Results for other families are in the Appendix L.

as in CharSpan, yields slightly better results.

Combining noise and BPE-dropout: We see that combining CSN with BPE-dropout gives the best-performing results.

Performance on Less Similar Languages: We evaluate the model's performance on languages that are less lexically similar to the considered languages and have different scripts. The languages are Gujarati (Guj), Punjabi (Pan), Bengali (Ben), Telugu (Tel), and Tamil (Tam). We first perform script-conversion of these languages to HRL by Kunchukuttan (2020)). The training setup is similar to the Indo-Aryan family. Table 2 shows that the ELRLs, which are lexically similar to HRLs, demonstrate a larger performance gain, while those with less lexical similarity show limited improvement. This suggests that the model's effectiveness is closely tied to the lexical similarity of the languages in CharSpan.

Impact of Cross-lingual Transfer: In this analysis, we investigate the encoded representations of the sentences to gain insights into how performance improves with char-span noise augmentation. We collected pooled last-layer representations of the encoder for HRL and LRLs across all parallel test examples using BPE, unigram char-noise (UCN), and the *CharSpan* models. We then calculated the average cosine similarity scores across the test set, presented in Table 3. Notably, the *CharSpan* model demonstrates high similarity, indicating a well-aligned embedding space for enhanced cross-lingual transfer.

Importance of Selecting Right HRLs: Table 4 presents an analysis of the impact of lexically diverse HRLs used for training. Results indicate that the CharSpan model demonstrates a performance gain when lexically similar HRLs were considered for noise injection. When the HRLs are less lexically similar, a degradation in performance is observed. These findings indicate the importance of using lexically similar HRLs.

Model	Hne	Mag	Mai	Npi	San
Training with Lexically Similar HRLs: Hin, Mar, Pan, Guj, Ben					
BPE	43.04	45.08	39.51	31.92	29.29
Char-span Noise	45.89	45.82	41.67	34.40	30.34
Training with Lexically less similar HRLs: Hin, Tel, Tam, Mal, Ora					
BPE	41.87	42.27	36.95	30.50	26.95
Char-span Noise	39.93	40.34	37.98	29.20	25.84

Table 4: Analysis experiment to show zero-shot chrF scores with lexically diverse HRLs. Due to computational constraints, we have considered 1 million parallel data for each HRL.

Imapct of small ELRL parallel Data: Here, we combined small ELRLs parallel data with the HRLs training data for BPE and CharSpan model. The results are presented in Table 14 in the appendix K. The additional data boosts both model performance, and CharSpan still outperforms the BPE model.

6 Conclusion

This study presents a simple yet effective novel character-span noise (CSN) argumentation model, CharSpan, to facilitate better cross-lingual transfer from HRLs to closely related ELRLs. The approach generalizes to closely related HRL-ELRL pairs from three typologically diverse language families. The proposed model consistently outperformed all the baselines. To the best of our knowledge, we are the first to apply noise augmentation for the NLG task. In the future, we will extend CharSpan to other NLP tasks, combine it with pre-trained models, and investigate noise augmentation in English-to-ELRL MT task.

Limitations

The current work addresses only transfer from related LRLs to English. It still remains to be investigated if noise augmentation is beneficial for translation from English to extremely low-resource languages. We assume that the related languages also use the same script or scripts that can be easily mapped/transliterated to each other. This method might not be effective for transfer between related languages that are written in very different scripts e.g. Hindi is written in the Devanagari script, while Sindhi is written in the Perso-Arabic script.

Ethics Statement

This work did not involve any new data collection and did not employ any annotators for data collection. We use publicly available datasets for experiments reported in this work. Some of these datasets originate from webcrawls and we do not make any explicit attempt to identify any biases in these datasets and use them as-is.

References

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.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? a study on POS tagging for non-standardized languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 40–54, Dubrovnik, Croatia. Association for Computational Linguistics.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1693–1703, Berlin, Germany. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Marta R. Costa-jussà, Carlos Escolano, and José A. R. Fonollosa. 2017. Byte-based neural machine translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 154–158, Copenhagen, Denmark. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(107):1–48.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Google-2018. The wordpiece algorithm in open source bert. https://github.com/google-research/bert/blob/master/tokenization.py#L335-L358. Retrieved on 11/01/2023.

Rohit Gupta, Laurent Besacier, Marc Dymetman, and Matthias Gallé. 2019. Character-based nmt with transformer. *CoRR, abs/1911.04997.*

Georg Heigold, Stalin Varanasi, Günter Neumann, and Josef van Genabith. 2018. How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse? In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 68–80, Boston, MA. Association for Machine Translation in the Americas.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020a. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020b. 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.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting language relatedness for low web-resource language model adaptation: An Indic languages study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1312–1323, Online. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Anoop Kunchukuttan, Mitesh Khapra, Gurneet Singh, and Pushpak Bhattacharyya. 2018. Leveraging orthographic similarity for multilingual neural transliteration. *Transactions of the Association for Computational Linguistics*, 6:303–316.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jindřich Libovický and Alexander Fraser. 2020. Towards reasonably-sized character-level transformer NMT by finetuning subword systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2572–2579, Online. Association for Computational Linguistics.

Jindřich Libovický, Helmut Schmid, and Alexander Fraser. 2022. Why don't people use character-level machine translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2470–2485, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

I. Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Third Workshop on Very Large Corpora*.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the*

Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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.

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.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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.

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.

Uri Shaham and Omer Levy. 2021. Neural machine translation without embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 181–186, Online. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *arXiv preprint arXiv:2201.03110*.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 90–96.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online. Association for Computational Linguistics.

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. 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 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Baselines

We compare the proposed model performance with the following strong baselines:

- **Vanilla NMT (BPE; Sennrich et al. (2016b)):** Neural Machine Translation model training with the standard BPE algorithm.

- **WordDropout (Sennrich et al., 2016a):** In this baseline, randomly selected words in the source/target sentence have their embeddings set to 0. We have selected 10% words in the source sentence as the noise augmentations are done in the source.

- **SubwordDropout:** It is a variant of WordDropout baseline where we drop the BPE tokens instead of words.

- **WordSwitchOut (Wang et al., 2018):** This baseline employs a data augmentation technique where random words in both the source and target sentences are replaced with randomly selected words from their respective vocabularies. We have utilized the officially released implementation with a 10% word replacement rate.

- **SubwordSwitchOut:** It is a variant of WordSwitchOut baseline where we use the BPE tokens instead of words.

- **Overlap BPE (OBPE; Patil et al. (2022)):** The approach modifies the BPE algorithm to encourage more shared tokens between high-resource and low-resource languages tokens in the vocabulary. This model required a monolingual dataset for ELRLs. We use a small monolingual dataset, based on availability, for the ELRLs. Earlier work applied OBPE for NLU tasks only - we are the first to investigate it for MT.

- **Soft Decoupled Encoding (SDE; (Wang et al., 2019)):** In the SDE approach, the authors have designed a framework that effectively decouples word embeddings into two interacting components: representing the spelling of words and capturing the latent meaning of words. This modeling technique has demonstrated its effectiveness in improving the performance of low-resource languages. In our study, we utilized the officially released implementation of SDE.

- **BPE-Dropout (Provilkov et al., 2020):** It utilizes the BPE algorithm to learn the vocabulary and sample different segmentations for input text during training (on-the-fly).

- **Unigram Character Noise (UCN; Aepli and Sennrich (2022)):** Inspired by the UCN model, we augment character-level noise (with all three operations) instead of char-span, the rest of the setup is similar to CharSpan.

- **BPE → Char-Span Noise:** In this ablation, we first learn vocabulary with clean HRLs. Subsequently, character-span noise is augmented into training data. This will demonstrate the significance of learning the BPE vocab with the noisy dataset.

- **Char-Span Noise + BPE-Dropout:** In this model, we train the BPE-Dropout model with char-span noise augmented HRLs training dataset.

B CharSpan Algorithm

Algorithm 1 shows the details of the character span noise augmentation presented in the main paper.

C Dataset Details and Statistics

The details of all the datasets used with CharSpan model are presented in Table 5.

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
```

Family	Code	Language	Script	Family	Subgrouping	Res.	Train	Dev	Test	Data Source
1	Hin	Hindi	Devanagari	Indo-European	Indo-Aryan	High	10M	1000	2390	Ramesh et al. (2022)
	Mar	Marathi	Devanagari	Indo-European	Indo-Aryan	High	3.6M	1000	2390	Ramesh et al. (2022)
	Bho	Bhojpuri	Devanagari	Indo-European	Indo-Aryan	Low	-	-	1012	FLORES-200
	Gom	Konkani	Devanagari	Indo-European	Indo-Aryan	Low	-	-	2000	ILCI[4]
	Hne	Chhattisgarhi	Devanagari	Indo-European	Indo-Aryan	Low	-	-	1012	FLORES-200
	San	Sanskrit	Devanagari	Indo-European	Indo-Aryan	Low	-	-	1012	FLORES-200
	Npi	Nepali	Devanagari	Indo-European	Indo-Aryan	Low	-	-	1012	FLORES-200
	Mai	Maithili	Devanagari	Indo-European	Indo-Aryan	Low	-	-	1012	FLORES-200
	Mag	Magahi	Devanagari	Indo-European	Indo-Aryan	Low	-	-	1012	FLORES-200
	Awa	Awadhi	Devanagari	Indo-European	Indo-Aryan	Low	-	-	1012	FLORES-200
2	Spa	Spanish	Latin	Indo-European	Romance	High	6.6M	670	1131	Rapp (2021)
	Pot	Portuguese	Latin	Indo-European	Romance	High	4.8M	681	1103	Rapp (2021)
	Cat	Catalan	Latin	Indo-European	Romance	Low	-	-	1012	FLORES-200
	Glg	Galician	Latin	Indo-European	Romance	Low	-	-	1012	FLORES-200
3	Ind	Indonesian	Latin	Austronesian	Malay-Polynesian	High	0.5M	2500	3000	OPUS[5]
	Zsm	Malay	Latin	Austronesian	Malay-Polynesian	High	0.3M	1500	2000	OPUS
	Jav	Javanese	Latin	Austronesian	Malay-Polynesian	Low	-	-	1012	FLORES-200
	Sun	Sundanese	Latin	Austronesian	Malay-Polynesian	High	-	-	1012	FLORES-200
Others	Pan	Panjabi	Gurmukhi	Indo-European	Indo-Aryan	Low	1M*	1000*	1012	FLORES-200
	Guj	Gujarati	Gujarati	Indo-European	Indo-Aryan	Low	1M*	1000*	1012	FLORES-200
	Ben	Bengali	Bengali	Indo-European	Indo-Aryan	High	1M*	1000*	1012	FLORES-200
	Tam	Tamil	Tamil	Indo-European	Indo-Aryan	Low	1M*	1000*	1012	FLORES-200
	Tel	Telugu	Dravidian	Indo-European	Indo-Aryan	Low	1M*	1000*	1012	FLORES-200
	Mal	Malayalam	Malayalam	Indo-European	Indo-Aryan	Low	1M*	1000*	1012	FLORES-200
	Ora	Oriya	Oriya	Indo-European	Indo-Aryan	Low	1M*	1000*	1012	FLORES-200

Table 5: Dataset details and Statistics. * are obtained from Ramesh et al. (2022)

D Model Training Details

We used the FairSeq library (Ott et al., 2019) to train proposed CharSpan and other baseline models. Training and implementation details are presented in Table 6. The best checkpoint was selected based on validation loss. The training time for the Indo-Aryan family of languages was approximately 8 hours; for the Romance languages it was approximately 7 hours, and for the Malay-Polynesian languages, it was less than 1 hour. Each language inference was completed within a time frame of less than 5 minutes. Due to computational limitations, the performance of the model was reported based on a single run. During the generation process, a batch size of 64 and a beam size of 5 were used, with the remaining parameters set to the default values provided by FairSeq. For data-pre-processing and script conversion for Indic languages, we use the Indic NLP library[6].

architecture	encoder-decoder (transformers)
# encoder layers	6
# decoder layers	6
# parameters	46,956,544 shared
learning rate (lr)	$5e^{-4}$
optimizer	adam
dropout rate	0.2
input size	210 tokens (both side)
epochs	15
tokens per batch	32768
clip-norm	1.0
lr scheduler	inverse sqrt
# GPUs	8
type of GPU	V100 Nvidia
generation batch size	64
beam size	5

Table 6: Model implementation and training details

E Performance Evaluation with BLEU, BLEURT and COMET Metrics

BLEU[7], BLEURT and COMET scores are reported in Table 7, 8 and 9, respectively. We observe the same trends as reported in the main paper for chrF[8].

F Performance on High Resource Languages

The high-resource language performances are presented in Table 10. It can be observed that, even with the inclusion of noise augmentation, the proposed model exhibits only a slight decrease in performance for HRLs.

G Candidate Alphabets

The candidate alphabets for noise augmentation are shown in Fig. 3. For the Indo-Aryan language family, the Devanagari alphabet is used, while the Latin alphabet is used for the Romance and Malay-Polynesian language families.

H Ablation Study and Different Experimental Setups

In order to ascertain the optimal configuration of the proposed model, a comprehensive set of experiments, numbering approximately 200, were conducted. A selection of the key evaluation scores from these experiments is illustrated in Table 13.

I Language Similarity Histogram

As depicted in Fig. 4, a similarity analysis in the form of a heatmap for the selected language families and languages is presented. The analysis shows that extremely low-resource languages (ELRLs) are closely related to high-resource languages (HRLs). The lexical similarity between languages was measured using character-level longest common subsequence ratio (LCSR) metric (Melamed, 1995). Additionally, the similar head map is presented for less similar languages in Fig. 5. These languages were used in the multiple analyses.

J Sample Translation Examples

A few sample translations from the proposed CharSpan model are shown in Fig. 6.

K Impact of Additional Small ELRLs parallel Data

Here, we combined small ELRL parallel data with the HRLs training data for BPE and CharSpan model. The results are presented in Table 14. The inclusion of additional data boosts both model performance, and CharSpan still outperforms the BPE model.

[6] https://github.com/anoopkunchukuttan/indic_nlp_library

[7] computed with SacreBLEU BLEU signature: *nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1*

[8] computed with SacreBLEU chrF signature: *nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1*

Models	Indo-Aryan								Romance		Malay-Polynesian		Average
	Gom	Bho	Hne	San	Npi	Mai	Mag	Awa	Cat	Glg	Jav	Sun	
BPE	4.36	10.62	15.76	3.43	4.36	9.36	16.7	15.6	5.23	22.99	5.74	6.02	10.01
WordDropout	4.62	11.21	15.71	4.11	5.47	9.96	16.76	16.31	6.19	22.26	5.90	6.02	10.37
SubwordDropout	4.57	9.99	14.47	3.93	5.25	9.08	15.53	16.03	5.85	20.72	4.78	4.93	09.59
WordSwitchOut	4.03	10.75	15.86	3.56	4.92	9.91	16.85	15.54	5.27	21.97	5.95	6.35	10.08
SubwordSwitchOut	4.13	10.56	15.93	3.76	4.49	9.69	16.61	16.69	5.19	23.82	6.02	6.01	10.24
OBPE	4.65	10.62	16.31	3.63	4.95	9.18	16.88	15.69	5.03	22.91	5.33	5.81	10.08
SDE	4.77	10.69	16.21	3.66	5.42	9.86	16.80	16.03	5.47	23.51	5.88	6.39	10.39
BPE-Dropout	5.24	11.33	15.64	3.71	4.94	10.00	16.62	16.63	5.94	24.07	5.79	6.65	10.54
unigram char-noise	5.21	12.62	18.29	3.81	6.55	11.29	19.47	18.95	11.82	24.09	7.35	6.87	12.19
BPE → CSN (our)	5.39	13.06	19.00	4.48	7.01	13.17	20.30	19.69	11.91	24.27	7.51	7.30	12.75
CHARSPAN (our)	5.77	13.01	19.52	4.63	7.13	13.43	20.81	20.36	12.21	24.72	7.52	7.32	13.03
CHARSPAN + BPE-Dropout (our)	**5.81**	**13.81**	**21.03**	**4.64**	**8.10**	**14.33**	**22.11**	**21.25**	**12.64**	**25.35**	**7.52**	**7.31**	**13.65**

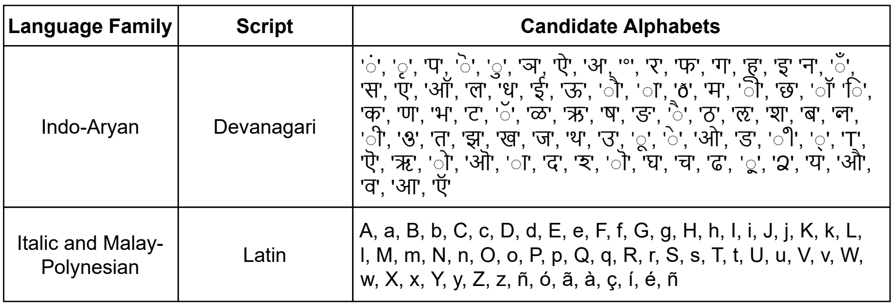
Table 7: Zero-shot BLEU scores results for ELRLs → English machine translation

Models	Indo-Aryan								Romance		Malay-Polynesian		Average
	Gom	Bho	Hne	San	Npi	Mai	Mag	Awa	Cat	Glg	Jav	Sun	
BPE	0.461	0.494	0.522	0.414	0.461	0.494	0.537	0.549	0.357	0.495	0.403	0.401	0.474
WordDropout	0.467	0.502	0.527	0.419	0.465	0.497	0.542	0.565	0.344	0.496	0.392	0.391	0.475
SubwordDropout	0.454	0.493	0.513	0.393	0.459	0.481	0.526	0.554	0.319	0.468	0.382	0.383	0.460
WordSwitchOut	0.456	0.501	0.528	0.395	0.445	0.497	0.552	0.551	0.309	0.477	0.381	0.381	0.464
SubwordSwitchOut	0.459	0.494	0.519	0.415	0.455	0.496	0.535	0.555	0.365	0.496	0.383	0.385	0.467
OBPE	0.466	0.496	0.518	0.419	0.459	0.491	0.537	0.551	0.431	0.428	0.396	0.381	0.464
SDE	0.486	0.499	0.515	0.511	0.496	0.542	0.543	0.553	0.440	0.481	0.406	0.405	0.489
BPE-Dropout	0.474	0.494	0.501	0.413	0.461	0.481	0.522	0.555	0.443	0.443	0.407	0.412	0.467
unigram char-noise	0.471	0.523	0.547	0.403	0.456	0.486	0.571	0.592	0.495	0.501	0.403	0.405	0.487
BPE → CSN (our)	0.469	0.528	0.553	0.400	0.459	0.491	0.579	0.595	0.499	0.511	0.405	0.413	0.491
CHARSPAN (our)	0.471	0.541	0.571	0.403	0.471	0.534	0.593	0.616	0.502	0.555	**0.419**	0.422	0.508
CHARSPAN + BPE-Dropout (our)	**0.478**	**0.548**	**0.582**	**0.421**	**0.478**	**0.535**	**0.604**	**0.623**	**0.505**	**0.567**	**0.419**	**0.429**	**0.515**

Table 8: Zero-shot BLEURT (computed with *BLEURT-20* checkpoint) scores results for ELRLs → English

Models	Indo-Aryan								Romance		Malay-Polynesian		Average
	Gom	Bho	Hne	San	Npi	Mai	Mag	Awa	Cat	Glg	Jav	Sun	
BPE	0.536	0.632	0.671	0.511	0.525	0.593	0.694	0.716	0.494	0.714	0.444	0.441	0.580
WordDropout	0.551	0.648	0.678	0.521	0.557	0.618	0.695	0.728	0.565	0.715	0.451	0.443	0.597
SubwordDropout	0.541	0.638	0.659	0.528	0.548	0.607	0.684	0.717	0.524	0.686	0.437	0.428	0.583
WordSwitchOut	0.544	0.647	0.681	0.522	0.563	0.621	0.706	0.719	0.529	0.702	0.453	0.452	0.594
SubwordSwitchOut	0.542	0.641	0.668	0.521	0.528	0.601	0.694	0.721	0.567	0.718	0.452	0.451	0.592
OBPE	0.541	0.629	0.667	0.504	0.527	0.589	0.691	0.715	0.492	0.721	0.363	0.611	0.587
SDE	0.549	0.636	0.666	0.513	0.529	0.591	0.697	0.735	0.513	0.731	0.357	0.618	0.594
BPE-Dropout	0.549	0.638	0.644	0.506	0.531	0.589	0.677	0.721	0.504	0.747	0.373	0.626	0.592
unigram char-noise	0.562	0.679	0.701	0.536	0.573	0.634	0.728	0.754	0.554	0.741	0.408	0.621	0.624
BPE → CSN (our)	0.557	0.676	0.706	0.542	0.581	0.651	0.724	0.755	0.561	0.751	0.403	0.622	0.627
CHARSPAN (our)	0.571	0.695	0.723	**0.556**	0.611	0.685	0.747	0.772	0.568	**0.759**	**0.417**	0.627	0.644
CHARSPAN + BPE-Dropout (our)	**0.579**	**0.705**	**0.733**	0.551	**0.616**	**0.687**	**0.757**	**0.778**	**0.572**	0.756	0.414	**0.631**	**0.648**

Table 9: Zero-shot COMET (computed with *Unbabel/wmt22-comet-da* model) scores results for ELRLs → English



Language Family	Script	Candidate Alphabets
Indo-Aryan	Devanagari	'ं', 'ृ', 'प', 'े', 'ु', 'ञ', 'ऐ', 'अ', '॰', 'र', 'फ', 'ग', 'ह', 'इ' 'न', 'ँ', 'स', 'ए', 'ऑ', 'ल', 'ध', 'ई', 'ऊ', 'ौ', 'ा', 'ઠ', 'म', 'ी', 'छ', 'ॉ' 'ि', 'क', 'ण', 'भ', 'ट', 'ॅ', 'ळ', 'ऋ', 'ष', 'ङ', 'ै', 'ठ', 'ऌ', 'श', 'ब', 'ॡ', 'ी', 'ઉ', 'त', 'झ', 'ख', 'ज', 'थ', 'उ', 'ू', 'े', 'ओ', 'ड', 'ी', '्', 'T', 'ऐ', 'ॠ', 'ो', 'ओ', 'ा', 'द', 'ર', 'ो', 'घ', 'च', 'ढ', 'ू', '૨', 'य', 'औ', 'व', 'आ', 'ऍ'
Italic and Malay-Polynesian	Latin	A, a, B, b, C, c, D, d, E, e, F, f, G, g, H, h, I, i, J, j, K, k, L, l, M, m, N, n, O, o, P, p, Q, q, R, r, S, s, T, t, U, u, V, v, W, w, X, x, Y, y, Z, z, ñ, ó, ã, à, ç, í, é, ñ

Figure 3: Candidate alphabets for noise augmentation, specifically for the insertion and substitution operations

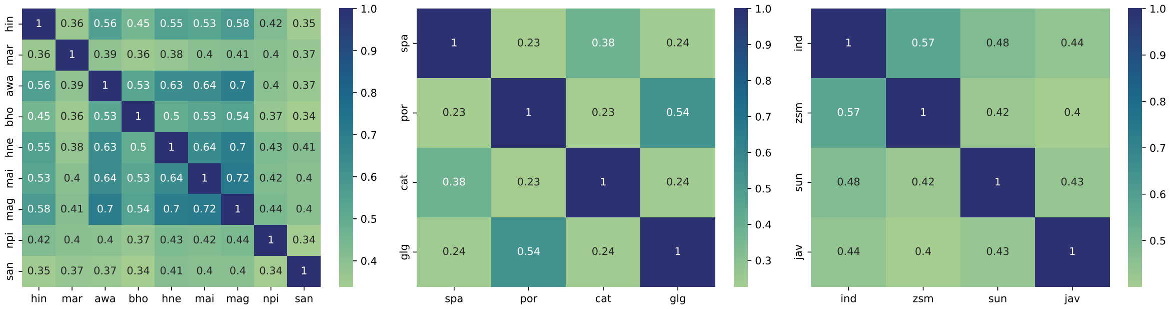


Figure 4: Lexical similarity (LCSR) heatmaps for three languages families. The Indo-Aryan languages are considered to use the Devanagari script while the Latin script is used by the other two language families.

XX → EN	Indo-Aryan				Romance				Malay-Polynesian			
Models	**BLEU**		**chrF**		**BLEU**		**chrF**		**BLEU**		**chrF**	
	Hin	**Mar**	**Hin**	**Mar**	**Spa**	**Pot**	**Spa**	**Pot**	**Ind**	**Zsm**	**Ind**	**Zsm**
BPE	37.44	26.31	64.04	54.47	41.44	35.38	68.71	63.27	29.61	21.76	58.31	49.14
WordDropout	36.54	26.31	63.27	53.96	39.32	32.73	66.89	60.86	27.59	20.42	56.72	48.22
SubwordDropout	36.64	26.22	63.46	54.57	39.84	33.04	67.56	61.58	26.73	18.80	57.02	48.82
WordSwitchOut	34.12	23.84	60.98	51.84	35.27	30.63	63.25	58.38	27.04	19.60	55.69	46.93
SubwordSwitchOut	37.11	26.03	63.78	54.06	42.26	35.68	68.65	62.97	27.12	19.76	55.72	47.34
OBPE	37.32	26.90	64.05	55.03	41.81	36.44	68.17	63.45	28.14	21.83	57.11	49.21
SDE	37.22	26.19	63.98	55.44	41.41	35.51	68.61	62.89	29.11	21.52	58.25	48.98
BPE-Dropout	37.22	26.93	64.11	55.31	41.88	36.72	68.06	63.79	30.39	22.54	59.33	50.17
unigram char-noise	37.05	26.95	63.81	54.83	39.83	32.91	67.62	61.24	28.79	22.01	57.65	49.91
BPE → CSN (our)	36.66	26.93	63.80	54.84	39.92	32.22	66.83	61.06	27.84	22.16	57.15	50.19
CHARSPAN (our)	36.68	26.70	63.87	54.59	40.04	32.36	66.95	61.03	27.84	21.87	56.75	49.58
CHARSPAN + BPE-Dropout (our)	37.62	27.10	64.15	55.03	41.21	33.64	66.90	61.39	28.91	22.26	57.99	50.59

Table 10: BLEU and chrF Scores: High resource language performance for all three language families

XX → EN	Indo-Aryan				Romance				Malay-Polynesian			
Models	**BLEURT**		**COMET**		**BLEURT**		**COMET**		**BLEURT**		**COMET**	
	Hin	**Mar**	**Hin**	**Mar**	**Spa**	**Pot**	**Spa**	**Pot**	**Ind**	**Zsm**	**Ind**	**Zsm**
BPE	0.775	0.726	0.891	0.857	0.769	0.720	0.871	0.830	0.687	0.561	0.821	0.701
WordDropout	0.774	0.725	0.891	0.854	0.755	0.701	0.86	0.814	0.681	0.555	0.815	0.693
SubwordDropout	0.773	0.725	0.889	0.854	0.757	0.691	0.861	0.806	0.672	0.548	0.803	0.683
WordSwitchOut	0.756	0.706	0.879	0.842	0.707	0.651	0.826	0.775	0.665	0.547	0.804	0.688
SubwordSwitchOut	0.776	0.724	0.892	0.855	0.771	0.721	0.872	0.833	0.663	0.548	0.801	0.687
OBPE	0.777	0.731	0.893	0.861	0.766	0.727	0.863	0.821	0.672	0.551	0.811	0.697
SDE	0.772	0.721	0.889	0.856	0.765	0.721	0.866	0.832	0.679	0.558	0.818	0.699
BPE-Dropout	0.773	0.727	0.891	0.858	0.772	0.7281	0.881	0.839	0.706	0.586	0.838	0.729
unigram char-noise	0.775	0.731	0.892	0.857	0.756	0.683	0.861	0.798	0.681	0.574	0.815	0.716
BPE → CSN (our)	0.773	0.728	0.891	0.857	0.755	0.685	0.861	0.801	0.685	0.581	0.821	0.724
CHARSPAN (our)	0.775	0.726	0.892	0.856	0.755	0.681	0.861	0.799	0.671	0.569	0.829	0.714
CHARSPAN + BPE-Dropout (our)	0.775	0.726	0.892	0.856	0.768	0.683	0.877	0.801	0.685	0.582	0.823	0.726

Table 11: BLEURT and COMET Scores: High resource language performance for all three language families

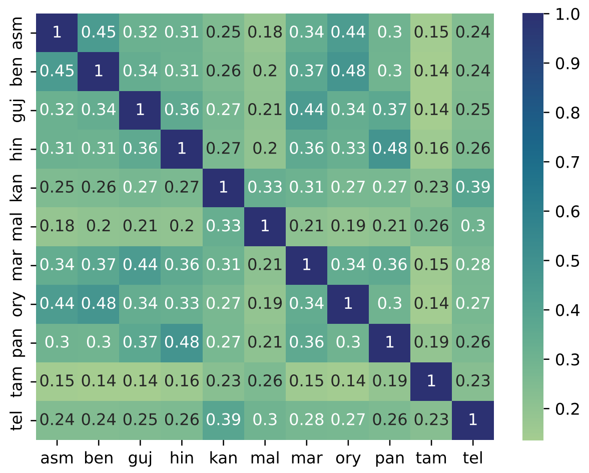


Figure 5: Lexical similarity heatmap for additional languages used in the analysis section. Here we have shown similarity scores for Assamese (asm), Bengali (ben), Gujrati (guj), Panjabi (pan), Hindi (him), Marathi (mar), Oriya (ory), Malayalam (mal), Kannada (kan), Tamil (tam) and Telugu (tel) languages.

L Effect of Cross-Lingual Transfer

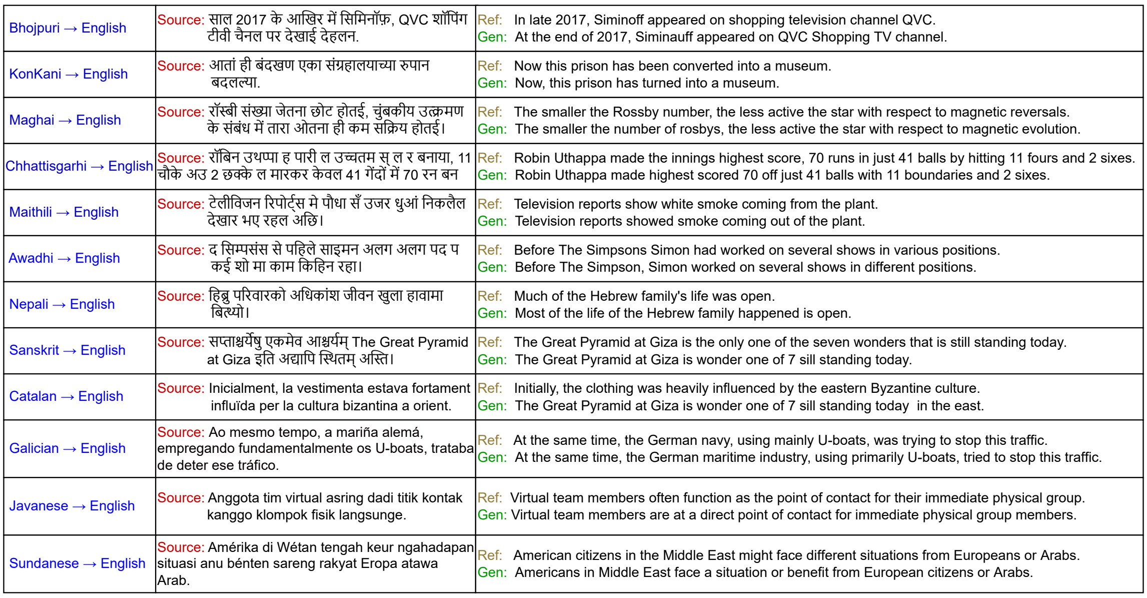
We did the following studies to understand why noise helps. The effectiveness of cross-lingual transfer depends on how well-aligned the representations of the HRL and ELRL are. Our hypothesis is that regularization with *char-level noise brings the representations of the HRL and ELRL closer to each other, thus improving cross-lingual transfer*. To measure these, we computed the cosine similarity of encoder representations from parallel HRL and ELRL sentences of 3 different models (baseline BPE, Unigram character-noise, CHARSPAN). The encoder representations were computed by mean-pooling the token representations of the top

Experimental Setup	Indo-Aryan							Average
	Bho	Hne	San	Npi	Mai	Mag	Awa	
ChrF Scores								
CHARSPAN with Hin, Mar, Pan, Guj, Ben	38.81	45.39	30.34	34.4	41.67	45.82	43.78	40.03
CHARSPAN with Hin, Mar, Pan, Guj	37.68	43.49	28.44	32.22	39.43	44.34	42.33	38.27
CHARSPAN with Hin, Mar, Pan	33.32	38.81	25.71	29.21	54.82	39.17	26.47	35.35
CHARSPAN with Hin, Mar	29.70	33.13	23.83	26.12	31.88	33.83	33.13	30.23
CHARSPAN with Hin	20.96	21.92	15.90	17.97	20.85	22.85	21.75	20.31
BLEU Scores								
CHARSPAN with Hin, Mar, Pan, Guj, Ben	10.46	15.97	4.87	7.02	11.83	16.32	14.65	11.58
CHARSPAN with Hin, Mar, Pan, Guj	9.55	14.32	3.92	5.99	9.85	14.71	13.47	10.25
CHARSPAN with Hin, Mar, Pan	7.41	10.21	2.91	4.63	7.88	11.01	9.89	7.70
CHARSPAN with Hin, Mar	5.30	7.06	2.40	3.20	5.00	7.28	6.96	5.31
CHARSPAN with Hin	2.03	2.27	0.6	0.97	1.77	2.23	2.39	1.75

Table 12: Zero-shot multilingual performance of char-span noise augmentation model. We have considered multiple combinations of high-resource languages for a multilingual setup. Due to computational constraints, 1 million parallel training data for each language was considered. All the languages are considered from the FLORES-200 test set.

Experimental Setups	BLEU (XX → EN)			chrF (XX → EN)		
	Gom	Bho	Hne	Gom	Bho	Hne
char-noise (9%-11% + replacement with only vowels)	4.77	11.21	15.17	28.08	40.36	46.13
char-noise (9%-11%+ replacement with only consonants)	4.79	11.25	15.3	26.95	40.51	46.17
char-noise (9%-11% + replacement with char sound similarity)	4.55	10.7	15.78	27.86	40.45	46.98
char-noise (9%-11% + with number and punctuation)	5.13	12.07	17.66	27.66	41.43	48.68
char-noise (9%-11% + only insertion)	5.04	12.3	17.81	27.50	41.87	48.74
char-noise (9%-11% + only replacement)	5.58	12.8	18.75	28.85	42.43	49.68
char-noise (9%-11%+ only deletion)	4.22	11.92	18.39	28.65	42.02	49.36
char-noise (4%-6% + all three operations + equal probability)	5.44	11.66	18.01	28.62	40.95	48.63
char-noise (14%-16% + all three operations + equal probability)	5.17	11.4	17.01	27.93	40.32	47.61
char-noise (9%-11% + all three operations + equal probability)	5.21	12.62	18.29	28.85	42.53	49.35
char-span noise (9%-11% + 1-3 grams + replacement: N random chars -> span)	3.80	8.80	13.11	25.38	28.22	43.39
char-span noise (9%-11% + 1-3 grams + insertion: 1 random chars -> span)	**5.84**	13.29	20.49	29.29	43.51	51.33
char-span noise (9%-11% + 1-3 grams + insertion: N random chars -> span)	4.81	12.21	17.36	26.98	41.26	47.91
char-span noise (9%-11% + 1-3 grams + all three operations + equal probability)	4.01	10.41	16.33	27.99	36.66	46.13
char-span noise (9%-11% + 1-2 grams + replacement and deletion + equal probability)	5.42	12.08	18.02	29.17	42.21	49.17
char-span noise (9%-11% + 1-4 grams + replacement and deletion + equal probability)	5.79	11.85	18.02	**29.71**	42.41	49.74
char-span noise (9%-11% + 1-5 grams + replacement and deletion + equal probability)	5.56	11.36	17.06	24.13	26.35	29.55
char-span noise (9%-11%+ 1-3 grams + replacement and deletion +unequal probability)	5.48	12.12	18.16	29.01	41.74	49.37
Proposed: char-span noise (9%-11% + 1-3 grams + replacement and deletion + equal probability)	**5.81**	**13.81**	**21.03**	**29.71**	**43.75**	**51.69**

Table 13: Ablation Study and Different Experimental Setups. Similar trends were observed for other ELRLs and language families.



Bhojpuri → English	Source: साल 2017 के आखिर में सिमिनॉफ़, QVC शॉपिंग टीवी चैनल पर देखाई देहलन.	Ref: In late 2017, Siminoff appeared on shopping television channel QVC. Gen: At the end of 2017, Siminauff appeared on QVC Shopping TV channel.
KonKani → English	Source: आतां ही बंदखण एका संग्रहालयाच्या रुपान बदलल्या.	Ref: Now this prison has been converted into a museum. Gen: Now, this prison has turned into a museum.
Maghai → English	Source: रॉस्बी संख्या जेतना छोट होतई, चुंबकीय उत्क्रमण के संबंध में तारा ओतना ही कम सक्रिय होतई।	Ref: The smaller the Rossby number, the less active the star with respect to magnetic reversals. Gen: The smaller the number of rosbys, the less active the star with respect to magnetic evolution.
Chhattisgarhi → English	Source: रॉबिन उथप्पा ह पारी ल उच्चतम स् ल र बनाया, 11 चौके अउ 2 छक्के ल मारकर केवल 41 गेंदों में 70 रन बन	Ref: Robin Uthappa made the innings highest score, 70 runs in just 41 balls by hitting 11 fours and 2 sixes. Gen: Robin Uthappa made highest scored 70 off just 41 balls with 11 boundaries and 2 sixes.
Maithili → English	Source: टेलीविजन रिपोर्ट्स मे पौधा सँ उजर धुआं निकलैत देखार भए रहल अछि।	Ref: Television reports show white smoke coming from the plant. Gen: Television reports showed smoke coming out of the plant.
Awadhi → English	Source: द सिम्पसंस से पहिले साइमन अलग अलग पद प कई शो मा काम किहिन रहा।	Ref: Before The Simpsons Simon had worked on several shows in various positions. Gen: Before The Simpson, Simon worked on several shows in different positions.
Nepali → English	Source: हिब्रु परिवारको अधिकांश जीवन खुला हावामा बित्थ्यो।	Ref: Much of the Hebrew family's life was open. Gen: Most of the life of the Hebrew family happened is open.
Sanskrit → English	Source: सप्ताश्चर्येषु एकमेव आश्चर्यम् The Great Pyramid at Giza इति अद्यापि स्थितम् अस्ति।	Ref: The Great Pyramid at Giza is the only one of the seven wonders that is still standing today. Gen: The Great Pyramid at Giza is wonder one of 7 sill standing today.
Catalan → English	Source: Inicialment, la vestimenta estava fortament influïda per la cultura bizantina a orient.	Ref: Initially, the clothing was heavily influenced by the eastern Byzantine culture. Gen: The Great Pyramid at Giza is wonder one of 7 sill standing today in the east.
Galician → English	Source: Ao mesmo tempo, a mariña alemá, empregando fundamentalmente os U-boats, trataba de deter ese tráfico.	Ref: At the same time, the German navy, using mainly U-boats, was trying to stop this traffic. Gen: At the same time, the German maritime industry, using primarily U-boats, tried to stop this traffic.
Javanese → English	Source: Anggota tim virtual asring dadi titik kontak kanggo klompok fisik langsunge.	Ref: Virtual team members often function as the point of contact for their immediate physical group. Gen: Virtual team members are at a direct point of contact for immediate physical group members.
Sundanese → English	Source: Amérika di Wétan tengah keur ngahadapan situasi anu bénten sareng rakyat Eropa atawa Arab.	Ref: American citizens in the Middle East might face different situations from Europeans or Arabs. Gen: Americans in Middle East face a situation or benefit from European citizens or Arabs.

Figure 6: Zero-shot Sample generations with CHARSPAN model for ELRLs.

Setup	Gom	Bho	Hne	San	Npi	Mai
BPE	26.75	39.75	46.57	27.97	30.84	39.79
BPE+ELRL$_{\text{par}}$	26.54	42.66	52.52	31.88	38.09	43.22
CSN	29.71	43.75	51.69	31.40	36.52	45.84
CSN+ELRL$_{\text{par}}$	29.65	45.39	53.38	33.92	39.66	47.18

Table 14: Translation quality (chrF) with an additional 1000 ELRL-English parallel sentences (ELRL$_{\text{par}}$).

layer of the encoder. The Table -15 shows the results (we report average results over the test set). We can clearly see that the similarity of encoder representations significantly increases in noise-augmented models. Further, CHARSPAN provides a good improvement over unigram char-noise which reflects improved translation quality.

M Error Analyses

M.1 Basline Generations are Transliterated

Fig. 7 presents a few sample examples where baseline models give generation error. Here, we look for transliteration errors. It can observed that many of the source words are directly transliterated in target generation for baseline models however the proposed CHARSPAN model successfully mitigate these error.

M.2 Grammatical Well-Formedness

It is often observed that the generations are grammatically not sound and such features are easily missed by the performance evaluation metrics like ChrF and BLEU. With this error analysis, we aim to investigate the grammatical well-formedness of generations from different baseline models. To score the grammatical well-formedness, we use L'AMBRE tool[9]. The results are reported in Table 16. For simplicity, we have shown results for only the Indo-Aryan family. The *CharSpan* shows better Grammatical formation than BPE and Unigram char-noise model across all ELRL.

These error analyses further provide evidence that the performance gains are truly genuine for the CHARSPAN model.

N Literature Review

In this section, we presented details of three threads of literature review related to the proposed work. This is summarized in Section 2 of the main paper.

N.1 MT for Low-resource Languages

Due to the unavailability of the large bi-text dataset for low-resource languages, much of the existing research focuses on *multilingual* MT. This enables cross-lingual transfer (Nguyen and Chiang, 2017; Zoph et al., 2016) and allows related languages to learn from each other (Fan et al., 2021; Costa-jussà et al., 2022; Siddhant et al., 2022). While this direction has gained significant attention, the performance improvement for LRLs as compared to HRLs has been limited (Tran et al., 2021) and remains an open area of research. In another thread, efforts have been made for MT models directly from the monolingual dataset (Artetxe et al., 2018; Lample et al., 2018; Lewis et al., 2020). These unsupervised approaches show promise, but still require a large amount of monolingual data, which should ideally match the domain of the HRLs (Marchisio et al., 2020). However, for many LRLs, monolingual datasets are not available (Artetxe et al., 2020). In contrast, we propose a model that does not require any bi-text/monolingual dataset and is scalable to any number of LRLs/dialects.

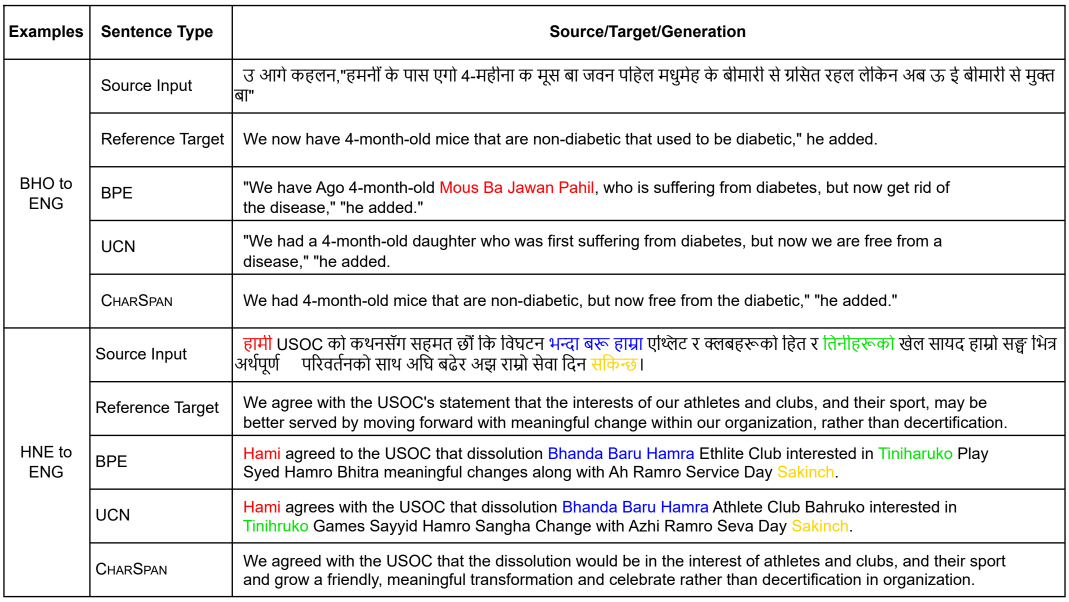
N.2 Vocabulary Adaptation for MT

Early exploration of character-based MT showed the promise (Chung et al., 2016; Lee et al., 2017) with coverage and robustness (Provilkov et al., 2020; Libovický and Fraser, 2020). However, recent modeling concludes a number of challenges (Gupta et al., 2019; Libovický and Fraser, 2020) in terms of training/inference times and performance as compared to the subwords models. Specifically, Shaham and Levy (2021) shows that character MT and Byte MT (Costa-jussà et al., 2017) have worse performance than the Byte Pair Encoding (BPE; (Sennrich et al., 2016b)) model and limits their practical usage (Libovický et al., 2022). The effectiveness of the BPE algorithm (Gage, 1994) is reported for NMT (Sennrich et al., 2016b) and serval other NLP tasks (Liu et al., 2019). Other algorithms like Sentencepiece (Kudo and Richardson, 2018) and Wordpiece (Google-2018)

[9] https://github.com/adithya7/lambre

Models	Indo-Aryan							Romance		Malay-Polynesian		Average
	Bho	Hne	San	Npi	Mai	Mag	Awa	Cat	Glg	Jav	Sun	
BPE	0.761	0.793	0.701	0.744	0.762	0.809	0.792	0.721	0.813	0.731	0.736	0.760
UCN	0.853	0.888	0.765	0.821	0.849	0.897	0.883	0.803	0.879	0.813	0.811	0.842
CHARSPAN	**0.871**	**0.909**	**0.789**	**0.858**	**0.868**	**0.913**	**0.901**	**0.831**	**0.903**	**0.846**	**0.856**	**0.867**

Table 15: Average cosine similarity between representations of source HRLs and source LRLs. UNC: Unigram Char-Noise



Examples	Sentence Type	Source/Target/Generation
BHO to ENG	Source Input	उ आगे कहलन,"हमनीं के पास एगो 4-महीना क मूस बा जवन पहिले 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.

Figure 7: The generation errors (transliteration) from different baseline models. The proposed CHARSPAN model successfully mitigates those errors. Colors indicate the corresponding transliteration in a generation.

Models	Indo-Aryan						
	Bho	Hne	San	Npi	Mai	Mag	Awa
BPE	0.9782	0.9813	0.9444	0.9624	0.9647	0.9784	0.9812
UCN	0.9754	0.9616	0.9504	0.9592	0.947	0.9708	0.9753
CHARSPAN	**0.9856**	**0.9865**	**0.9658**	**0.9735**	**0.9802**	**0.9842**	**0.9836**

Table 16: Grammatical Well-Formedness for different models with L'AMBRE

are similar to BPE. We take inspiration from existing works and proposed a model on BPE.

Given the potential of the BPE model, various methodologies have been developed for vocabulary modification/generation/adaption (Provilkov et al., 2020; Khemchandani et al., 2021; Patil et al., 2022; Minixhofer et al., 2022). In particular, the work of Provilkov et al. (2020) utilizes the BPE algorithm to generate the vocabulary and sample different segmentations during training. Patil et al. (2022) introduce an extension of BPE, referred to as Overlapped BPE (OBPE), which takes into account both HRLs and LRLs tokens during vocabulary creation. They demonstrate the effectiveness of this approach in only NLU tasks. In contrast, in this study, we adopt the standard BPE model on noisy HRL data for the MT task.

N.3 Surface/Lexical Level Noise for MT

Several previous studies (Sperber et al., 2017; Koehn and Knowles, 2017; Karpukhin et al., 2019; Vaibhav et al., 2019) have examined the use of noise augmentation strategies, including substitution, deletion, insertion, flip, and swap, at various levels of text granularity for machine translation. These strategies are explored to stabilize/improve the robustness of the model with naturally occurring noises, such as spelling mistakes. Further, these noising schemes are utilized to obtain non-canonical text in adversarial settings (Heigold et al., 2018). Close to ours, Aepli and Sennrich (2022) proposed a character-based noise model to transfer the supervision from HRLs to LRLs in a zero-shot setting. They evaluated the proposed model on two NLU tasks with the pre-trained model. Unlike this, we have trained the model from scratch for the machine translation task which is very different and more challenging than NLU tasks. Moreover, we explore the *span-denoise* which outperformed char denoise-based models and emerged as a desirable MT model for extremely low-resource languages and dialects.