The University of Edinburgh's Submission to the WMT22 Code-Mixing Shared Task (MixMT)

Faheem Kirefu, Vivek Iyer, Pinzhen Chen, Laurie Burchell
School of Informatics, University of Edinburgh
{fkirefu, vivek.iyer, pinzhen.chen, laurie.burchell}@ed.ac.uk

SUBTASK 1 (Hi + En -> Hg) | **SUBTASK 2 (Hg -> En)**

1. DATASETS

Dataset	Generation Method	Pair(s)
HinGe (Srivastava and Singh, 2021)	Provided by organizers	Hi -> Hg En -> Hg
L3Cube-HingCorpus (Nayak and Joshi, 2022)	Hg->En + Hg->Hi BT by XLM model	Hi (BT) -> Hg En (BT) -> Hg
CC100-Hindi Romanized (Conneau et al., 2020)	Hg->En + Hg->Hi BT by XLM model	Hi (BT) -> Hg En (BT) -> Hg
Transliterated Samanantar (Ramesh et al., 2021)	Transliteration of Hi->Hg using AI4Bharat	En -> Hg (Hi Transl.) Hi -> Hg (Hi Transl.)

Dataset	Generation Method	Pair(s)
PHINC (Srivastava and Singh, 2020)	Provided by organizers	Hg->En
ToxicWiki	Toxic content filtered from WikiMatrix	Hg->En
Sentiment140 (Sahni et al., 2017)	Public domain dataset	Hg->En
Transliterated Samanantar (Ramesh et al., 2021)	Transliteration of Hi->Hg using AI4Bharat	Hg (Hi Transl.) -> En

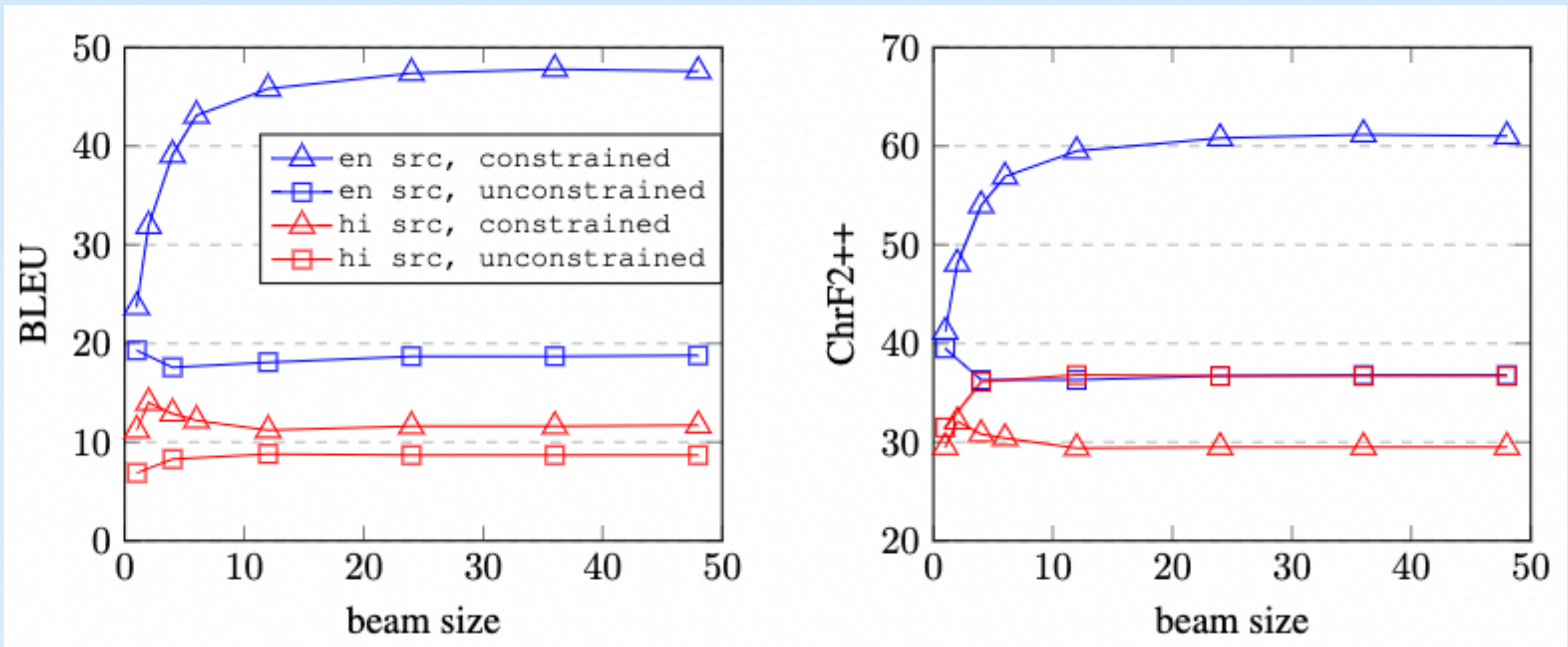
2. EXPERIMENTS

Subtask 1

- Training paradigm: 1) General domain Training (BT + transliterated corpora), followed by 2) Fine-tuning on HinGE dataset
- We explore Constrained Decoding to constrain Code-switched sentences to Hindi and English inputs

Approach	BLEU	ChrF++	TER	WER
Baseline (Unconstrained)	**18.1**	**44.0**	**64.5**	85.7
Constrained Decoding	15.0	38.7	73.6	**57.0**

- Constrained Decoding underperforms as generated Hg output closely resembles En src sentences, likely due to noise in Hg references, whereas Unconstrained produces more diverse translations.



Subtask 2

- Training paradigm: 1) General domain Training (BT + transliterated corpora), followed by 2) Fine-tuning on Sentiment140 + ToxicWiki, and lastly 3) Fine-tuning on PHINC
- Final model was an ensemble of 4 Hg-En models

Approach	BLEU	ChrF++	TER	WER
Single model	24.5	47.0	65.1	72.0
Ensemble (of 4)	**25.5**	**48.7**	**62.9**	**70.5**

- We also explored another pretraining paradigm: Aligned Augmentation.
- Though it resolved some spelling issues and grammatical inconsistencies over random baselines, it did not improve over our original ensemble models.
- Likely reasons include: high-resource setting leading to forgetting, noise in social media test data (vs pretraining data), usage of non-Hindi languages etc.

Approach	BLEU	ChrF++	TER	WER
Random	24.3	45.2	68.4	**74.6**
Aligned Augmentation	**24.4**	**46.2**	**68.2**	74.9

3. FINAL RESULTS

	BLEU	ChrF++	TER	WER	ROUGE-L	Human eval
UEdin (Subtask-1)	26.9	52.7	55.2	56.2	57.9	3.85

	BLEU	ChrF++	TER	WER	ROUGE-L	Human eval
UEdin (Subtask-2)	28.7	51.2	59.1	61.3	62.5	3.75

Performance Highlights!

✓ UEdin ranked 2nd on the automatic leaderboards for both subtasks
✓ In human evaluation, tied for 1st in Subtask 1 and 2nd in Subtask 2
✓ Our results are comparable to the top-performing system(s)