Synthetic Data Generation & Joint Learning for Robust Code-Mixed Translation

Kartik, Sanjana Soni, Anoop Kunchukuttan, Tanmoy Chakraborty, & Md. Shad Akhtar

COLING LREC 2024

LCS2
LABORATORY FOR COMPUTATIONAL SOCIAL SYSTEMS

What is Code-mixing / Code-switching ?

- **Code-mixing** is mixing of two or more languages in a single utterance.
  - Commonly seen in multilingual communities, e.g., **Hindi**-**English**, **Spanish**-**English**, Cantonese-Sanghaiese etc.
- **CM Machine Translation:** Translation of codemix text to another non-codemix language.

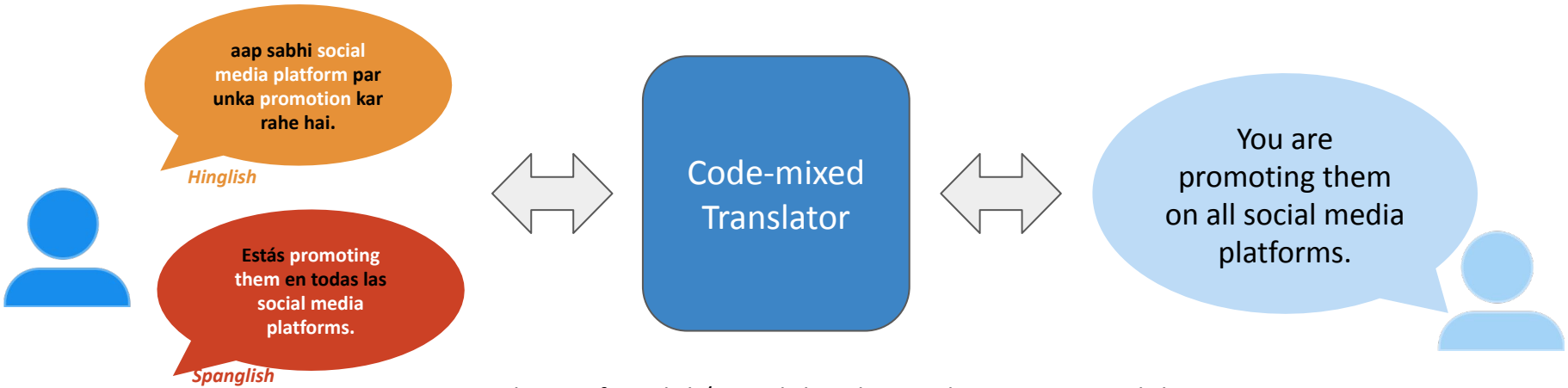
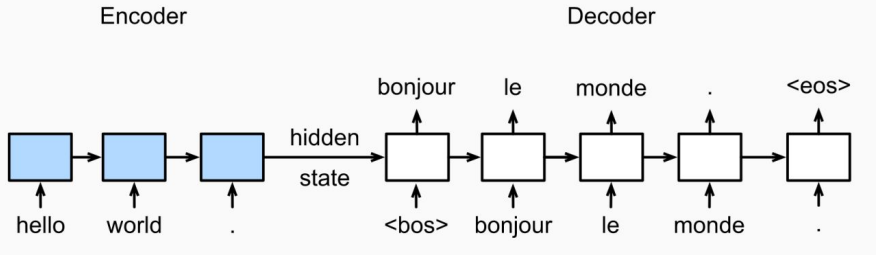


Fig: Translation of Hinglish/Spanglish code-mixed sentence to English.

Challenges with Neural Machine Translation

- NMT models (eg. enc-dec, BART, T5) need a vast amount of parallel data for satisfactory performance

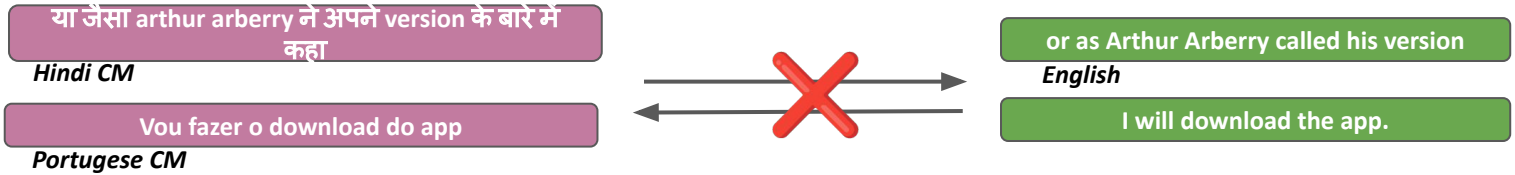


- NMT models are brittle to even a slight amount of input noise.

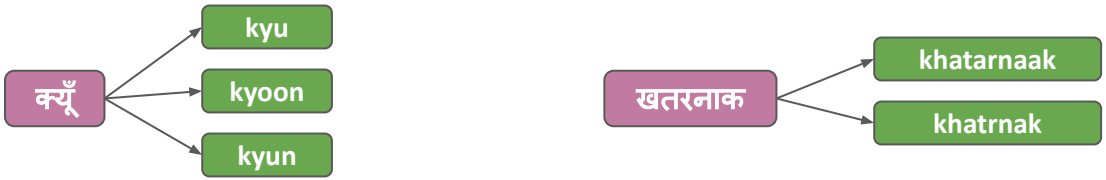
  (Belinkov and Bisk, 2017, Synthetic and Natural Noise Both Break Neural Machine Translation)

Challenges with Codemix Translation

- Unavailability of **parallel codemix data**; Online Social Media the only source of CM text



- **No standardization:** High susceptibility to spelling variations, typographical errors and misspellings



Contributions

- Pipeline for generating synthetic codemix parallel corpus from non-CM corpora
- **HINMIX**: the first open-source large-scale Hinglish Code-Mixed parallel corpus consisting of ~4.2M parallel sentences, with a gold dev and test set.
- Robust Code-Mixed Translation (**RCMT**) encoder-decoder model to handle real-world noisy codemix input.
- Zero-Shot CMT from **Bengali** CM to English without any parallel bengali CM corpus.

CM Data Generation pipeline

- Decide matrix language (Hindi) and embedded language (English)
- Obtain POS tags for each word & shortlist words such that POS tags belong the inclusion list (nouns, identifiers, adjectives).
- Build substitution dictionary from a parallel sentence using a trained word-level alignment model (fastalign).
- Substitute *r* words at random to obtain multiple sentences and filter using perplexity.
- Transliterate and add word-level perturbations like switching, omission, typos and shuffles.

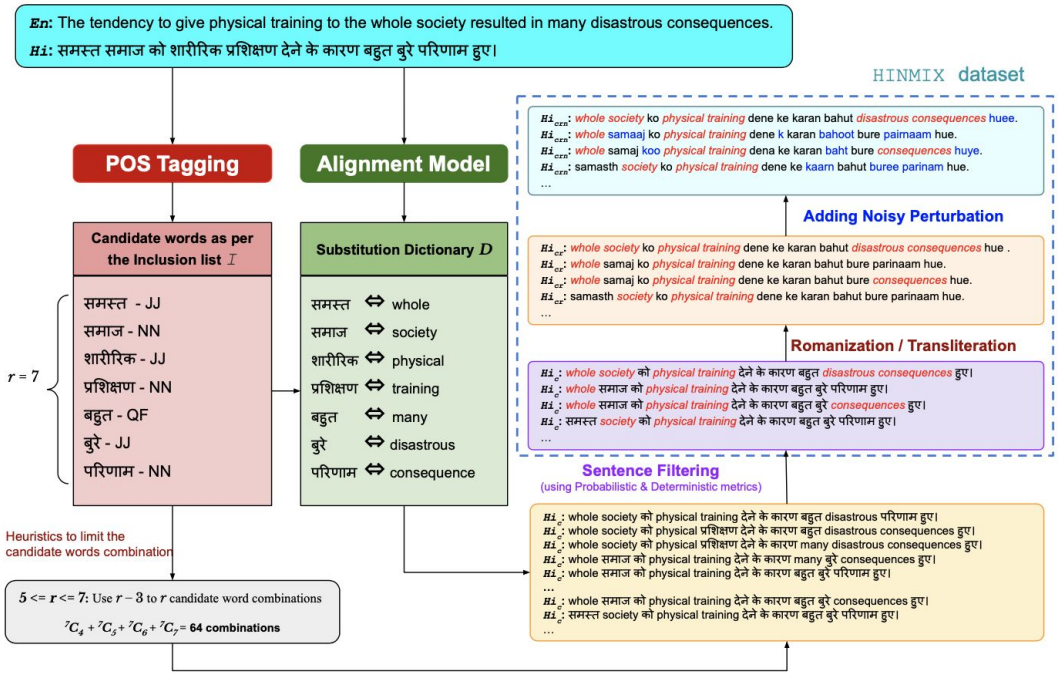


Figure 1: Process of code-mixed sentence generation in HINMIX.

HINMIX Test Data Guidelines

- *Gold-standard Hinglish-English dataset* created by **two** bilingual Hindi-English annotators
  - Annotators aged 25-35, Hindi as first language with English as second language
  - Given Hindi sentences, annotators asked to come up with Hinglish conversions as a first thought in the mind within 5 seconds
- No standardized romanization scheme used for Hindi words
  - Annotators transliterated based on personal understanding of word structure and sound pattern
  - No fixed spelling of any word results in natural noise to ensure robustness

Examples from HINMIX test set (Hi_cr - En)

Hinglish	English
jabki congress is baat par agree nahin ho saki ki aage ki proceedings karni hai or nahin bahut se state wait nahin kar rahe.	And while Congress can't agree on whether to proceed, several states are not waiting.
yah hamare country ke liye reality men mandatory thing hai.	This really is a must for our nation.
Police ke according, uski death dubne se hui hai.	According to the police his death was due to drowning.

HINMIX Dataset Statistics

- Train set contains **4.2M** pairs in 5 Hi forms; Development set has **280**, and Test set includes **2507** CM parallel sentence pairs.
- CodeMix Index (CMI) measures the percentage of code-mixing in a sentence; Switch Point Fraction (SPF) calculates the complexity of code-mixing (as % of switching b/w languages) in a sentence.
- Dev/Test have higher complexity and code-mixing percentage.
- Test set sentences are also much longer on average than the train set.

Statistics	Type	Sentence-level				Token-level					Char-level	
		#Sent	#Unique	CMI	SPF	# Hi_{src}	#En_{src}	#EN_{tgt}	Mean	Median	Mean	Median
Train	Synthetic	4.2M	0.67M	27.9	44.3	0.25M	0.11M	0.19M	100.9	88	18.24	16
Dev	Gold	280	280	32.6	47	711	667	1392	65.6	64	12.17	12
Test	Gold	2507	2507	32.4	45.5	4194	5923	11255	124.9	111	22.8	20

Table 2: Statistics of HINMIX code-mixed dataset.

Robust Codemix Translation (RCMT)

- *RCMT_roman:* We jointly train a **single encoder-decoder transformer** model in **three** directions: bidirectional Hindi-English code-mixed romanized corpus (Hicr⇌En) and Hindi to English **noisy** code-mixed romanized corpus (Hicrn →En)
- A proxy token is added to the source sentence to indicate the target language for decoding.
- Joint model is trained to optimize the sum of categorical cross-entropy (CE) loss with label smoothing across all pairs.
- Noisy data helps model become robust to spelling variations.

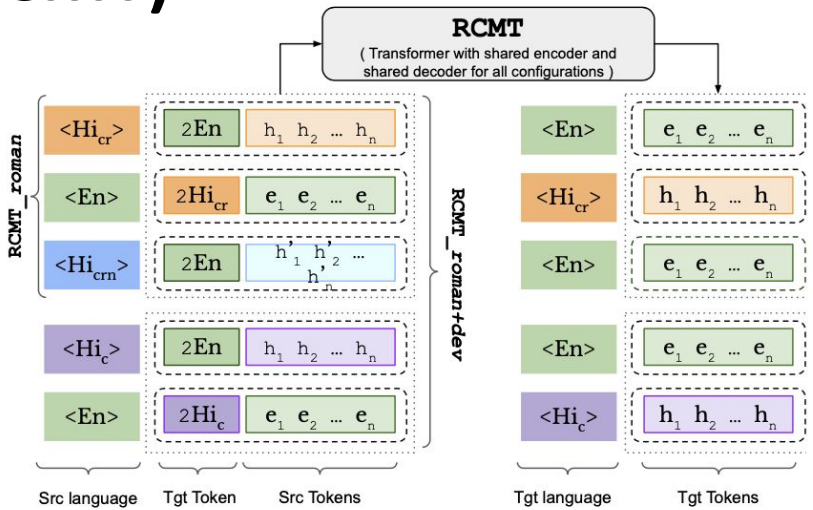


Figure 2: The proposed RCMT model. The subscripts c, r, and n denote codemix, romanized, and noisy version of a dataset. The target token $[2T]$ in the encoder input indicates the intended target language T followed by tokens in the source language S. The target tokens are passed to the decoder sequentially for model training.

Experiments & Results: Model Comparison

- RCMT handles **all-inclusive CM input** (Devanagari, English, romanized, and noisy words) and outperforms other baselines significantly
- Improvement over TFM but minor decline in results with the increase in the corpus/language. (*RCMT_roman → RCMT_roman+devan*)
- Transfer learning on Multilingual models does not work well.
- Noise robust MT architectures like MTT, AdvSR prove to be resilient to synthetic noise.

Model	c		c+r		c+r+n	
	B	M	B	M	B	M
TFM (Vaswani et al., 2017)	9.97	39.7	10.02	36.2	9.70	37.4
FCN (Gehring et al., 2017)	7.89	33.2	8.07	33.1	5.69	27.5
mT5 (Xue et al., 2021)	4.27	22.6	4.28	25.9	2.80	19.5
mBART (Liu et al., 2020b)	5.38	29.5	7.07	35.7	3.19	21.7
PtrGen (Gupta et al., 2020)	6.51	27.18	4.68	21.15	3.04	16.1
MTT (Zhou et al., 2019)	-	-	-	-	10.44	38.0
MTNT (Vaibhav et al., 2019)	-	-	8.48	35.1	5.92	28.0
AdvSR (Park et al., 2020)	-	-	9.63	36.7	7.28	32.7
RCMT_roman	-	-	13.58	**45.7**	**11.54**	**41.5**
RCMT_roman+devan	**13.81**	**46.2**	**13.72**	**45.7**	11.30	40.8

***Metrics*: Bleu (B) & Meteor (M)**

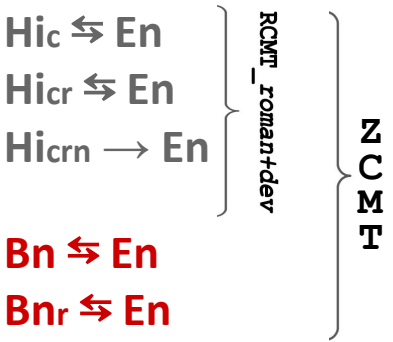
Experiments & Results: Dataset Comparison

- **Generalizability:** Joint learning shows good performance on non-codemix datasets as well.
- SpokenTutorial includes utterances transcribed from engineering, programming courses.
- RCMT proves robustness on real-world noisy tweets dataset - LinCE.
- Superior scores on HINMIX test demonstrates handling of longer and much complex sentences.

Datasets	RCMT_*roman*		RCMT_*roman+devan*	
	B	M	B	M
IITB (non-CM)	12.25	40.8	12.75	40.9
SpokenTutorial (CM)	22.58	52.1	23.07	52.5
LinCE (CM)	11.06	33.9	10.28	33.5
HINMIX (CM)	13.58	45.7	13.72	45.7

Experiments & Results : Zero Shot

- **Zero Shot Capabilities:** We train RCMT with Bangla-English (Bn-En) and Hinglish-English (Hi_{cr}-En) parallel corpora without building any CM corpus for Bangla.
- **Hypothesis**: Trained model (ZCMT) would be able to transfer the code-mixing behaviour onto the network activations in a zero-shot way
- Adding languages from the **same family** (Indo-Aryan) can sometimes improve the code-mixed translation quality despite varying scripts (Devanagari vs. Eastern-Nagari).



Model		Hindi		Bangla	
		B	M	B	M
MMT	–	13.59	45.0	**15.66**	47.7
	r	13.05	44.1	13.83	44.3
ZCMT	c	**14.00**	**46.7**	15.41	**49.8**
	c + r	13.69	46.1	14.01	47.6

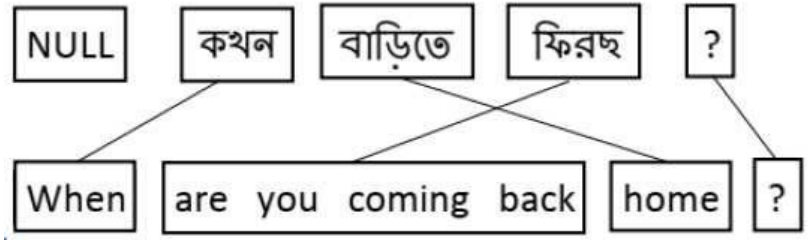
Limitations & Qualitative Analysis

- **Translation semantics:** Due to the alignment method, model has learned to copy words from the source sentence, missing a more suitable word.

Source ($\texttt{Hi}_{\texttt{cr}}$):	Is **thought** ko sabhi places par support nahin mila.
Target (En):	The **concept** is not a universal hit.
RCMT_*roman*	This thought did not support at all the places.
Source ($\texttt{Hi}_{\texttt{cr}}$)	Yah aapke relatives aur loved ones ke liye ek **complete** gift hai.
Target (En)	It is **perfect** gift for your relatives and loved ones.
RCMT_*roman*	This is a complete gift for your relatives and loved ones

Table 6: Sample translation of code-mixed ($\texttt{Hi}_{\texttt{cr}}$) sentences to English (En) produced by the proposed RCMT_*roman* model.

- **POS Tagging Errors:** In case a word in the source sentence is incorrectly POS tagged, its substitute word will not be appropriate.
- **Alignment Errors**: Incorrect word mapping between source-target completely alters the meaning



Thank you!

Code and Datasets can be found in this QR code

Questions?

***Contact*:**

kartikaggarwal98@gmail.com

sanjana19097@iiitd.ac.in

LCS2
LABORATORY FOR COMPUTATIONAL SOCIAL SYSTEMS