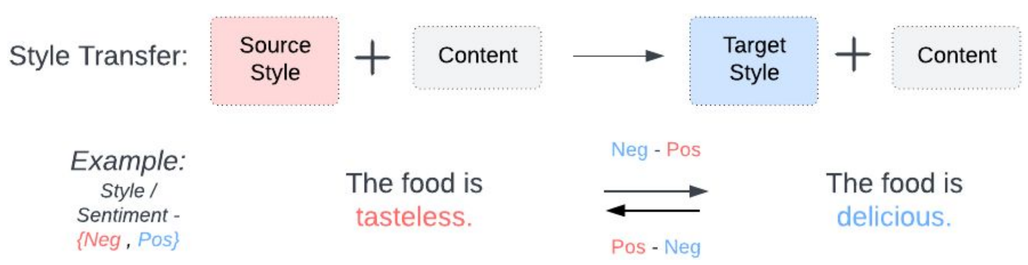
Multilingual Text Style Transfer: Datasets & Models for Indian Languages

Sourabrata Mukherjee mukherjee@ufal.mff.cuni.cz
Atul Kr. Ojha atulkumar.ojha@insight-centre.org
Akanksha Bansal akanksha.bansal@panlingua.co.in
Deepak Alok deepak.alok@panlingua.co.in
John P. McCrae john.mccrae @insight-centre.org
Ondřej Dušek odusek@ufal.mff.cuni.cz

Overview

- **Parallel datasets for sentiment transfer** (style transfer) **in 8 languages**
  - expanding our previous work on English & Bangla
- **Benchmark model experiments**, analysis & insights
- Showing significance of parallel data in style transfer



Details

- 1000+1000 positive+negative style-parallel sentences
- Hindi, Magahi, Malayalam, Marathi, Odia, Punjabi, Urdu, Telugu
- Benchmark models:
  - Parallel (finetuned mBART), Cross-lingual, Joint multilingual
  - Non-parallel:
    - Auto-encoder (AE) & Back-translation (BT) reconstruction
    - Masked Style Filling (MSF)
  - LLMs: Llama2, GPT-3.5
- Evaluation:
  - Style Transfer Accuracy: Sentiment Classifier
  - Content Preservation: BLEU, embedding similarity (CS)

Data Creation

- English Yelp restaurant reviews (Li et al. 2018), revised in our previous work

en: i will be going back and enjoying this great place !
hi: मैं वापस जाऊँगी और इस उम्दा जगह का आनंद लूँगी।
mag: हम फिर से जइबई आउ इ बढ़ीयाँ जगह के मजा लेबई!
ml: ഞാൻ തിരികെ പോയി ഈ മഹത്തായ സ്ഥലം ആസ്വദിക്കും!
mr: मी परत जाईन आणि या महान जागेचा आनंद घेईन !
or: ମୁଁ ଫେରିଯିବି ଏବଂ ଏହି ମହାନ ସ୍ଥାନକୁ ଉପଭୋଗକରିବି!
pa: ਮੈਂ ਵਾਪਸ ਜਾਵਾਂਗਾ ਅਤੇ ਇਸ ਵਧੀਆ ਸਥਾਨ ਦਾ ਆਨੰਦ ਮਾਣਾਂਗਾ!
ur: میں واپس جاؤں گا اور اس عظیم جگہ سے لطف اندوز ہوں گا!
te: నేను వెనక్కు వెళ్ళబోతున్నాను మరియు ఈ గొప్ప ప్రాంతాన్ని ఆనందిస్తాను.

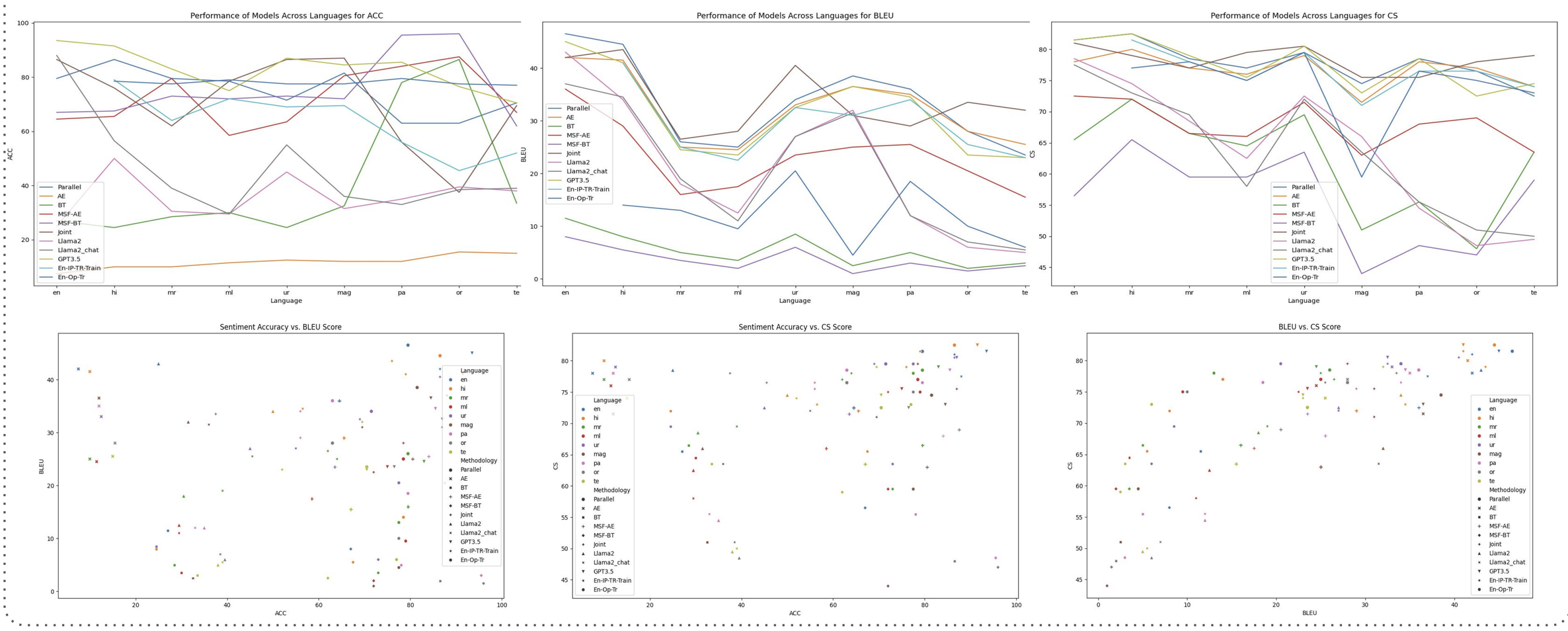
- native translator & validator for each language
- issues: ambiguity, cultural references, language differences

i won't be going back and suffering at this terrible place !
मैं इस भयानक जगह पर वापस जाकर पीड़ित नहीं होऊँगी!
हम फिर से नऽ जइबई आउ इ खराब जगह में कस्ट सहबई!
ഈ ഭയാനകമായ സ്ഥലത്ത് ഞാൻ തിരികെ പോയി കഷ്ടപ്പെടില്ല!
मी परत जाणार नाही आणि या भयानक ठिकाणी यातना सहन करणार नाही !
ମୁଁ ଆଉ ଏହି ଭୟଙ୍କର ସ୍ଥାନରେ କଷ୍ଟ ଭୋଗିବି ନାହିଁ!
ਮੈਂ ਵਾਪਸ ਨਹੀਂ ਜਾਵਾਂਗਾ ਅਤੇ ਇਸ ਬੇਕਾਰ ਜਗ੍ਹਾ 'ਤੇ ਦੁਖੀ ਨਹੀਂ ਹੋਵਾਂਗਾ!
میں واپس نہیں جاؤں گا اور اس خوفناک جگہ پر تکلیف نہیں دوں گا!
నేను వెనక్కి వెళ్ళి ఈ భయంకరమైన స్థలంలో బాధపడను}&

Experimental Results

- Parallel: balanced performance
- Non-parallel: generally underperform parallel
- Cross-lingual: competitive results
- Joint: strong, esp. English, Malayalam, Telugu & Urdu

- GPT-3.5 leads, but smaller & open models trained on little data can tie it
- Lower accuracy in lower-resource languages: more challenging



Code: https://github.com/souro/multilingual_tst
Data: https://github.com/panlingua/multilingual-tst-datasets

Presented at INLG 2024, Tokyo, Japan.

Funded by the European Union (ERC, NG-NLG, 101039303) and Charles University project SVV 260 968. Using resources of LINDAT/CLARIAH-CZ (Czech MEYS LM2018101).