Making Sentence Embeddings Robust to User-Generated Content

Lydia Nishimwe
Inria, France
lydia.nishimwe@inria.fr

Benoît Sagot
Inria, France
benoit.sagot@inria.fr

Rachel Bawden
Inria, France
rachel.bawden@inria.fr

User-Generated Content (UGC)

Ergographic phenomena (encoding simplification)

i **don wanna fyt witchu**

al b an **our l8**

c u 2moro

Transverse phenomena

i **aint playin**

idk

afaik

N. E. V. E. R

Marks of expressiveness

supe**rrrr !!!!**

<3

☺

!d10t

sh*t

Neologisms

The math is not **mathing.**

burkini

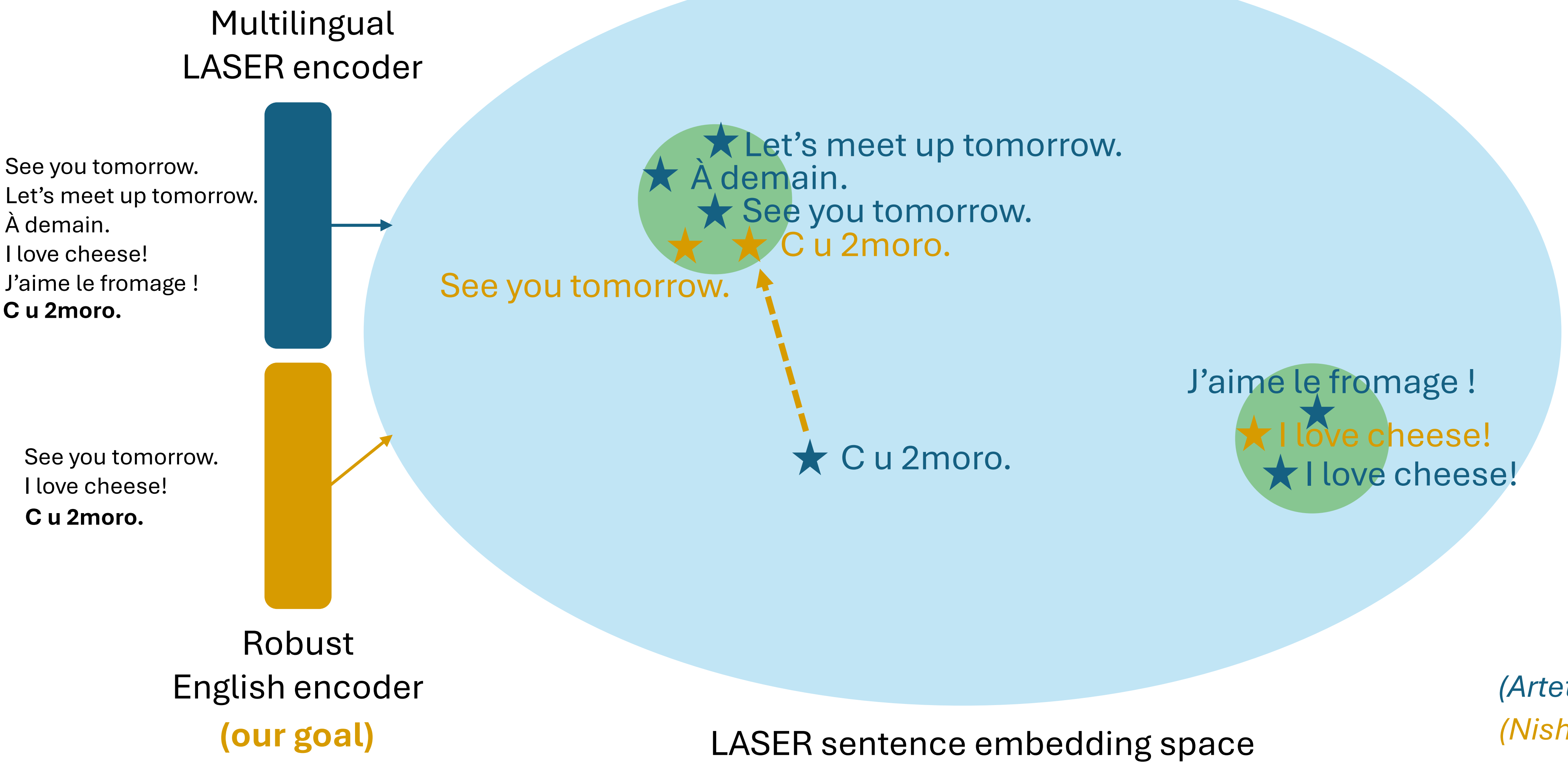
Foreign language influence

Cette fête a l'air **fun, let's go !**

likez et commentez

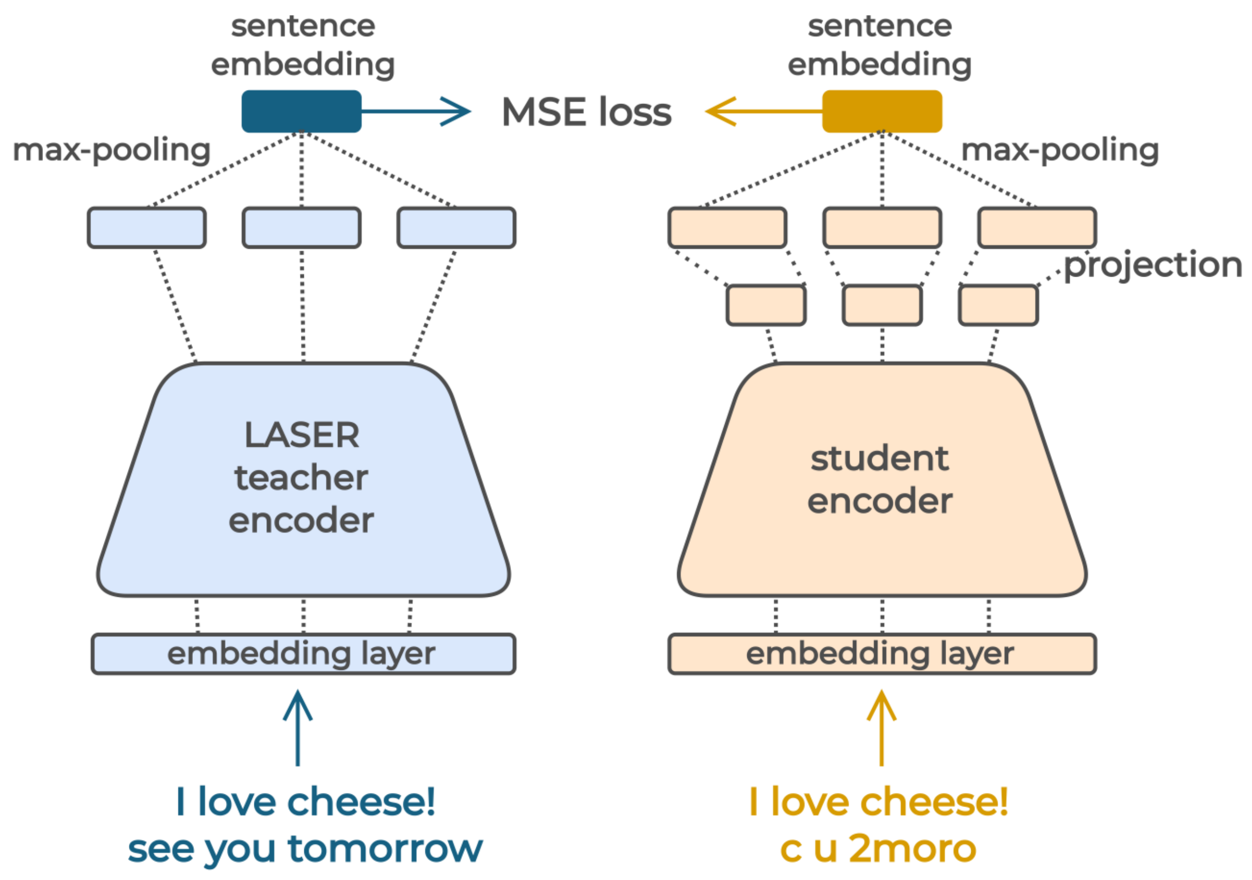
(Seddah et al., 2012)
(Zalmout et al., 2019)
(Sanguinetti et al., 2020)

Multilingual sentence embeddings



(Artetxe and Schwenk, 2019)
(Nishimwe et al., 2024)

Proposed approach: Teacher-Student training



- LASER (teacher):
  - 45M parameters
  - 5-layer bi-LSTM
  - 1024 output dimension
  - fixed during training
- RoLASER [Robust LASER] (student):
  - 108M parameters
  - 12-layer Transformer
  - 768 output dimension
  - projection layer -> 1024
- c-RoLASER (student):
  - 104M parameters
  - same as RoLASER, except for
  - Character-CNN input embedding layer

(Reimers and Gurevych, 2020; Duquenne et al., 2022; Mao and Nakagawa, 2023)

Generating artificial UGC (NL-Augmenter)

abbreviations, acronyms, slang

abr1 because → cuz

abr2 easy → ez

abr3 ASAP ↔ as soon as possible

slng jewellery → bling bling

contractions and expansions

cont I am ↔ I'm

week Monday ↔ Mon.

visual and segmentation

leet love → l0V3

spac hello there → h elloth ere

misspellings

fing tried → triwd

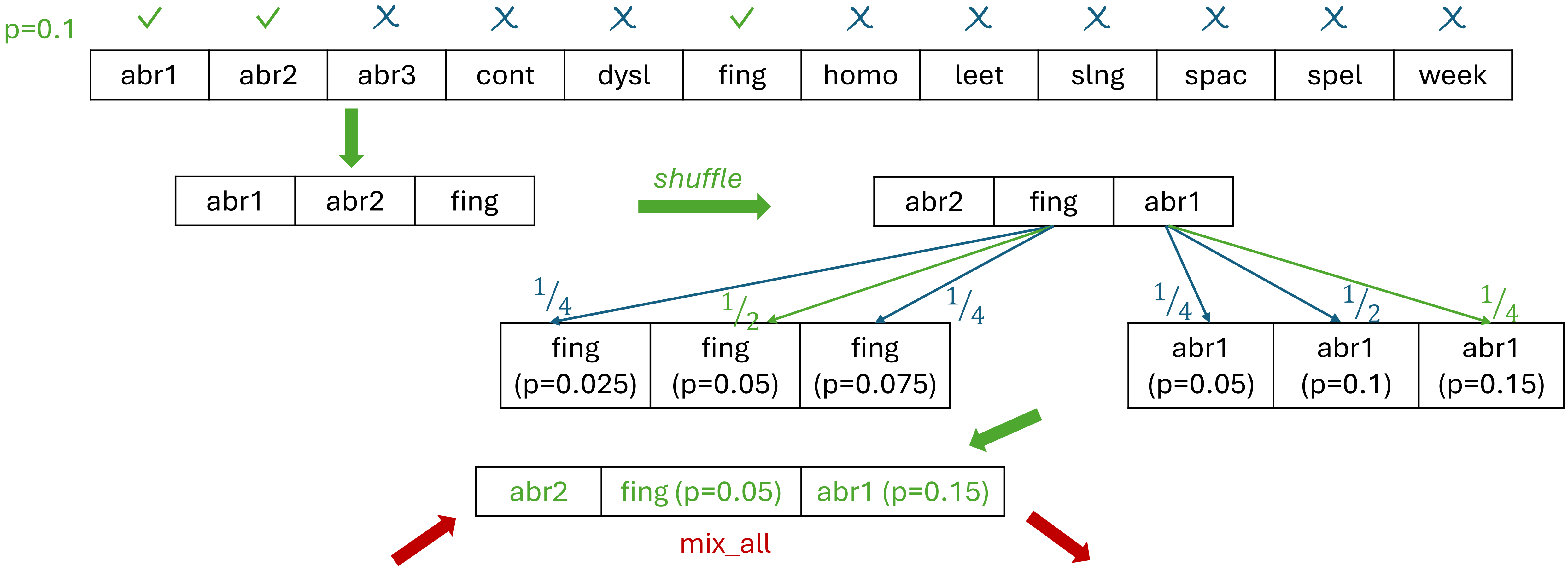
homo there ↔ their

dysl lose ↔ loose

spel absent → apsent

(Dhole et al., 2021)

Generating artificial UGC training data



"Luckily **nothing** happened **to** me, but I saw a macabre scene, as **people tried to** break windows in order **to get** out."

"Luckily **nthing** happened **2** me, but I saw a macabre scene, as **ppl triwd 2** break windows in order **2 gt** out."

Experimental setup

- **Training data:**
  - 2M “bilingual” standard-UGC sentences
  - 2M standard English sentences from the OSCAR dataset *(Ortiz Suárez et al., 2019)*
  - augmented with the *mix_all* transformation
- **RoLASER training:**
  - initialised with RoBERTa *(Liu et al., 2019)*
  - 98 epochs
- **c-RoLASER training:**
  - initialised with CharacterBERT *(El Boukkouri et al., 2020)*
  - 32 epochs

Evaluation data and metrics

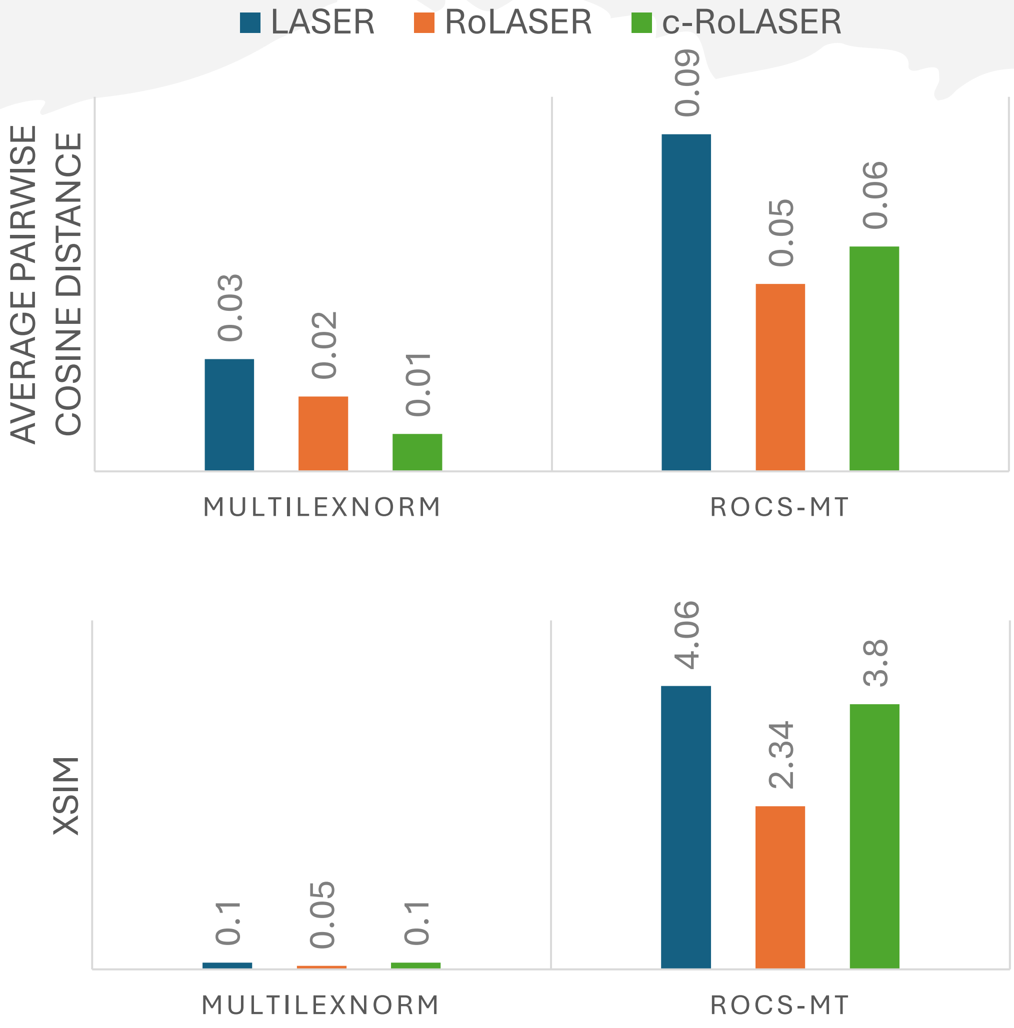
Data

- MultiLexNorm *(van der Goot et al., 2021)*
  - Twitter
  - 1967 standard ↔ UGC sentences in English
- RoCS-MT *(Bawden and Sagot, 2023)*
  - Reddit
  - 1922 standard ↔ UGC sentences in English
  - translations into other 5 languages
- FLORES-200 *(NLLB Team et al., 2022)*
  - WikiNews, WikiBooks, WikiVoyage
  - parallel texts in 200 languages
  - 997 dev and 1012 test sentences
  - artificially augmented

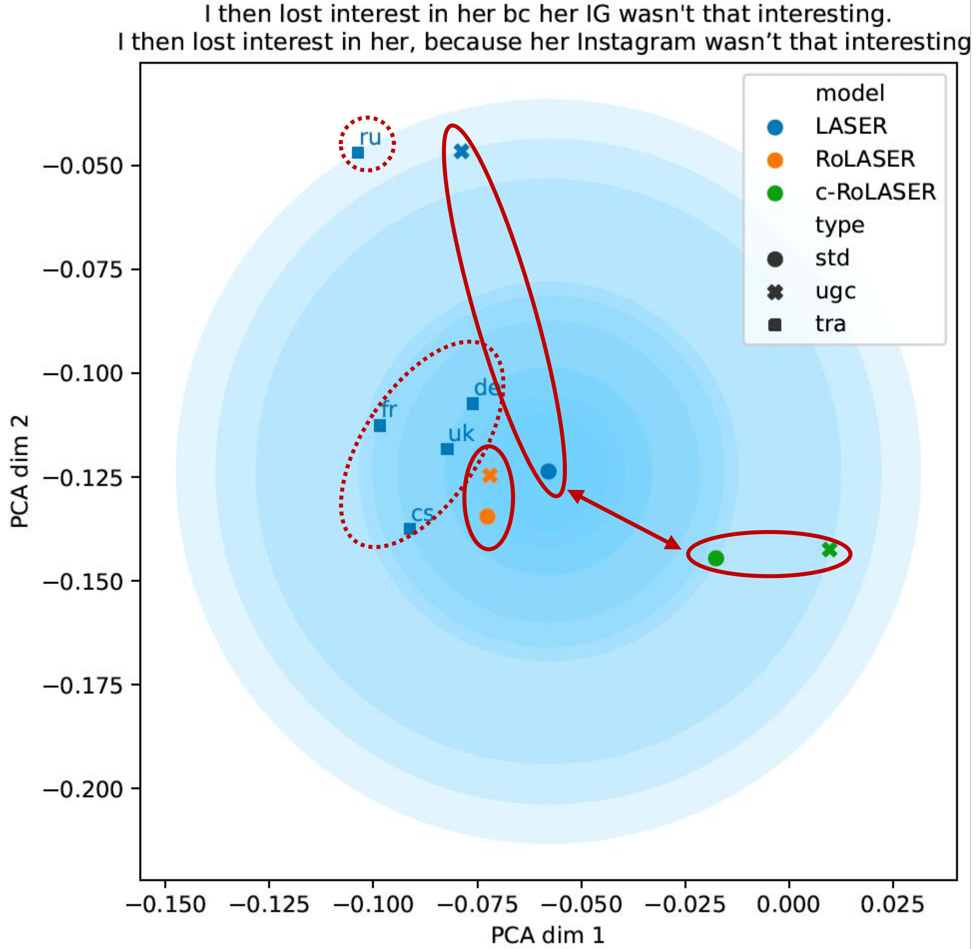
Metrics

- Average pairwise cosine distance
- xSIM *(Artetxe and Schwenk, 2019)*
  - cross-lingual similarity search
  - proxy metric for bitext mining
  - error rate of aligning translation pairs
- xSIM++ *(Chen et al., 2023)*
  - augmenting the English set of FLORES-200
  - altering the meaning
  - minimal surface changes
  - more challenging than xSIM

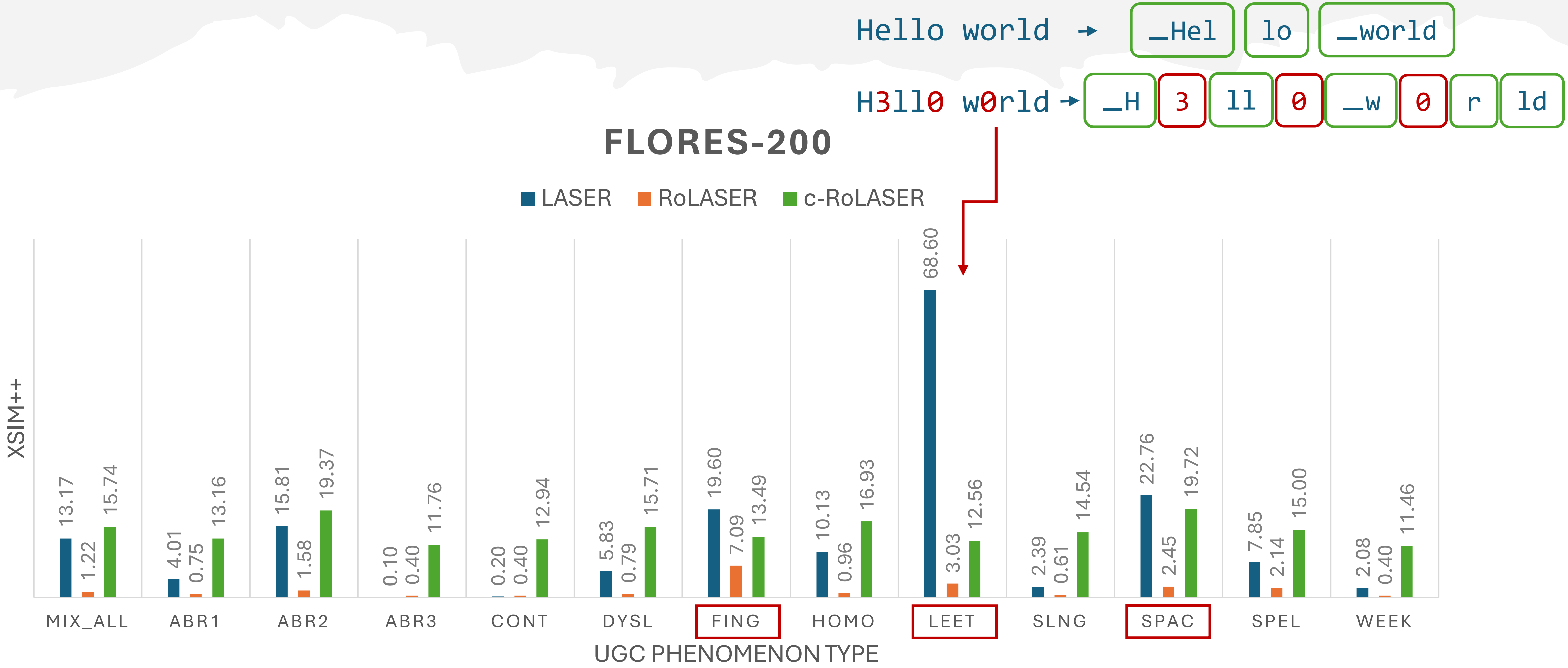
Evaluation on natural UGC



(lower is better)

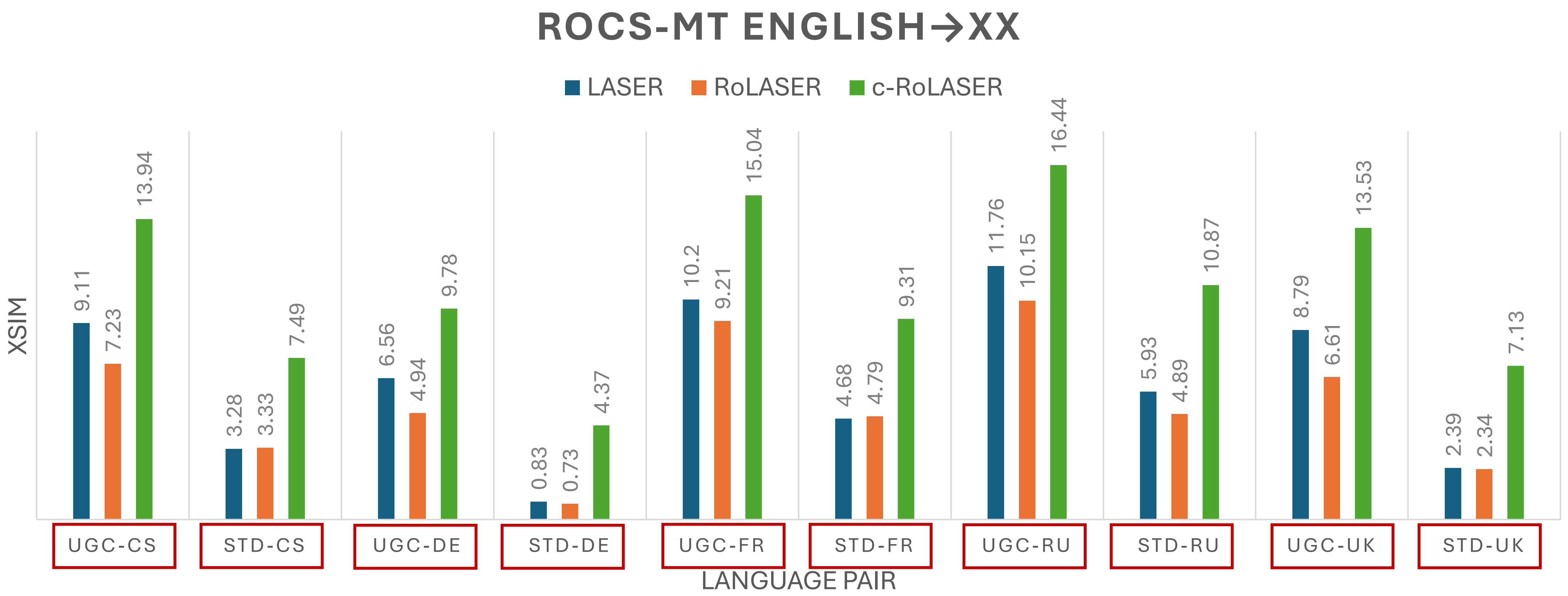


Evaluation on artificial UGC



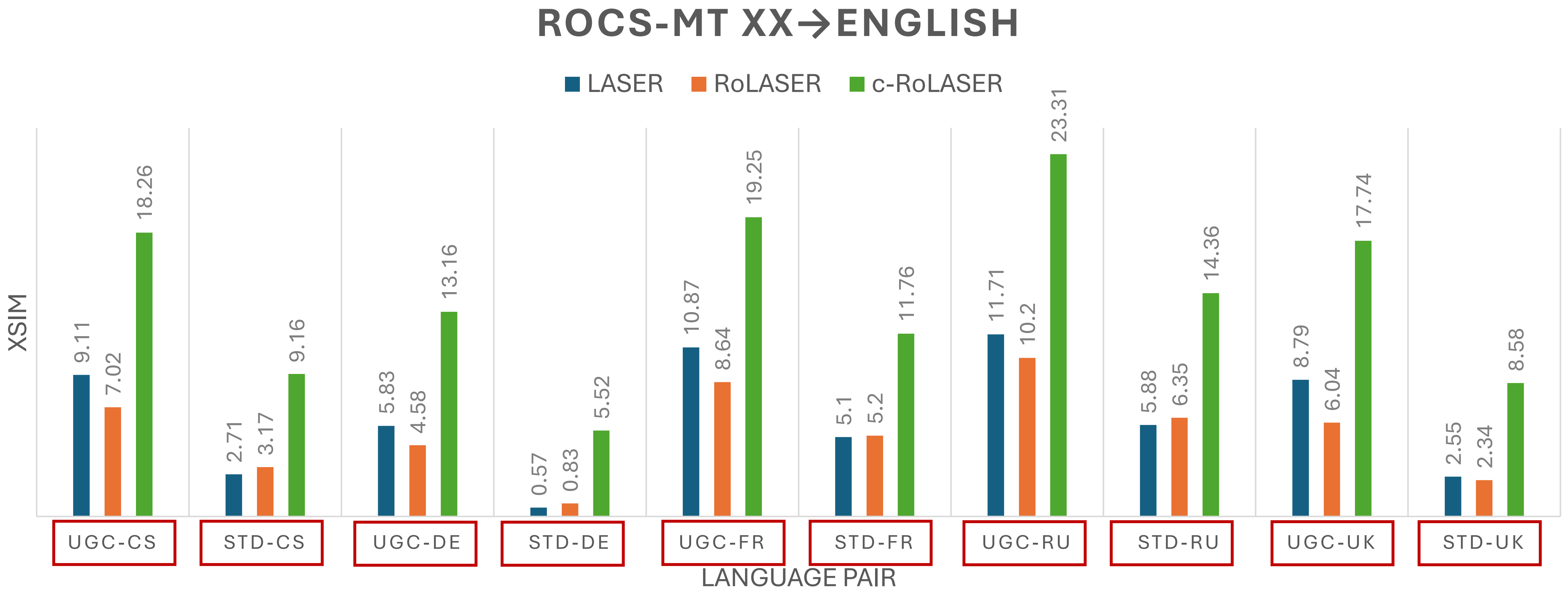
(lower is better)

Evaluation on UGC and standard data in a multilingual setting (1)



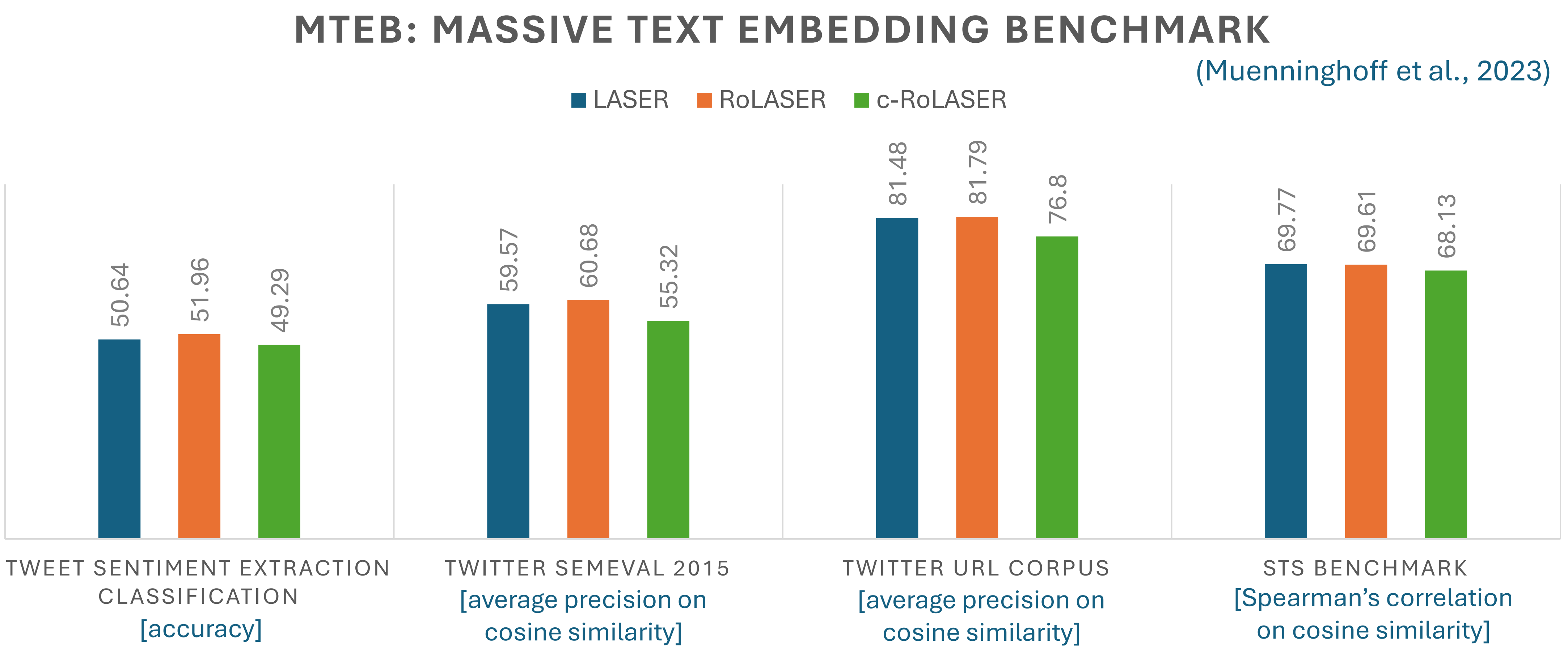
(lower is better)

Evaluation on UGC and standard data in a multilingual setting (2)



(lower is better)

Evaluation on downstream tasks



(higher is better)

Takeaways

Extending RoLASER to **more languages** and their corresponding UGC phenomena…

Future work

Approach: **Making LASER more robust to UGC English**

1. Teacher-Student training
2. Minimising the standard-UGC distance in the embedding space
3. Generating and training on synthetic UGC-like data

Results: **RoLASER is significantly more robust than LASER**

- on natural and artificial UGC
- on standard data and downstream tasks (improves/matches LASER's performance)

Findings:

1. c-RoLASER struggles to map its standard embeddings to LASER's
2. Most challenging UGC phenomena: character-level perturbations that shatter subword tokenisation

Read our paper for more details and references!