Making Sentence Embeddings Robust to User-Generated Content

Lydia Nishimwe, Benoît Sagot, Rachel Bawden
Inria, France
{lydia.nishimwe, benoit.sagot, rachel.bawden}@inria.fr

PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

Takeaways

Observation

LASER has **poor sentence representations** for **user-generated content**

Goal

Make LASER more robust by **reducing the standard-UGC distance** in the space

Results

RoLASER:

- ✓ is **significantly more robust than LASER** to user-generated content
- ✓ matches/**improves LASER's performance** on standard text

User-Generated Content (UGC)

Ergographic phenomena

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

superrrr !!!! **sh*t**

<3 ☺ **!d10t**

Neologisms

The math is not **mathing**.

burkini

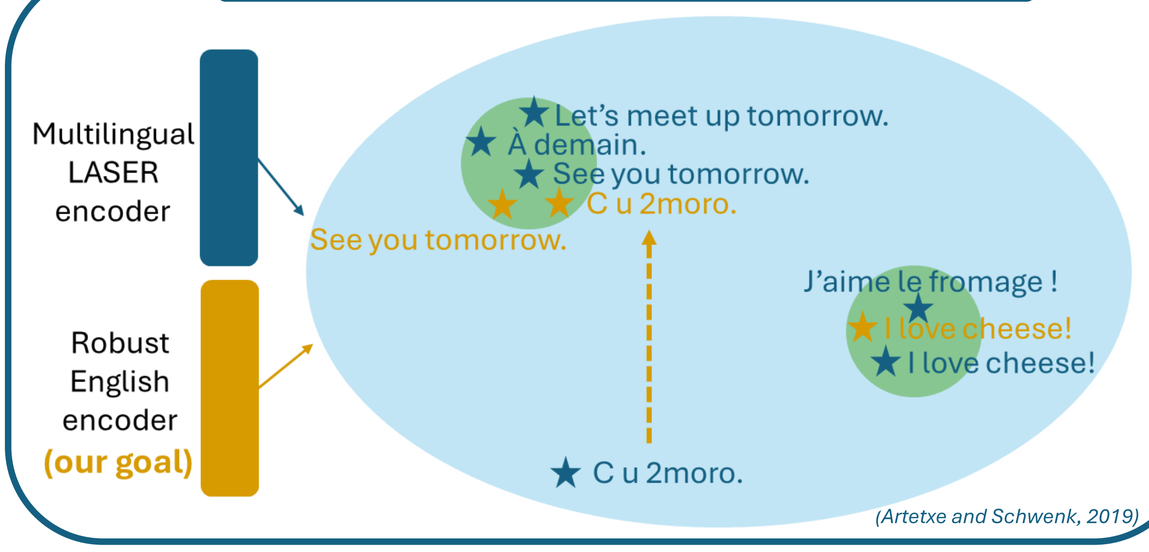
Foreign language influence

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

likez et commentez

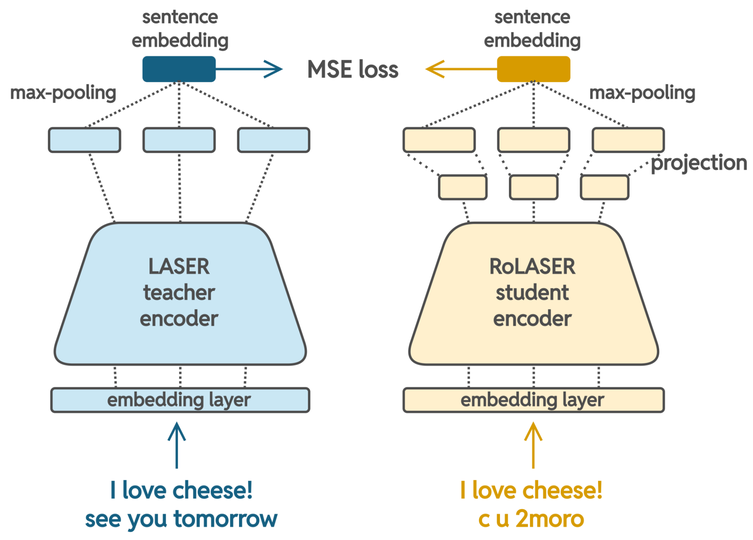
(Seddah et al., 2012; Zalmout et al., 2019; Sanguinetti et al., 2020)

LASER Sentence Embedding Space



(Artetxe and Schwenk, 2019)

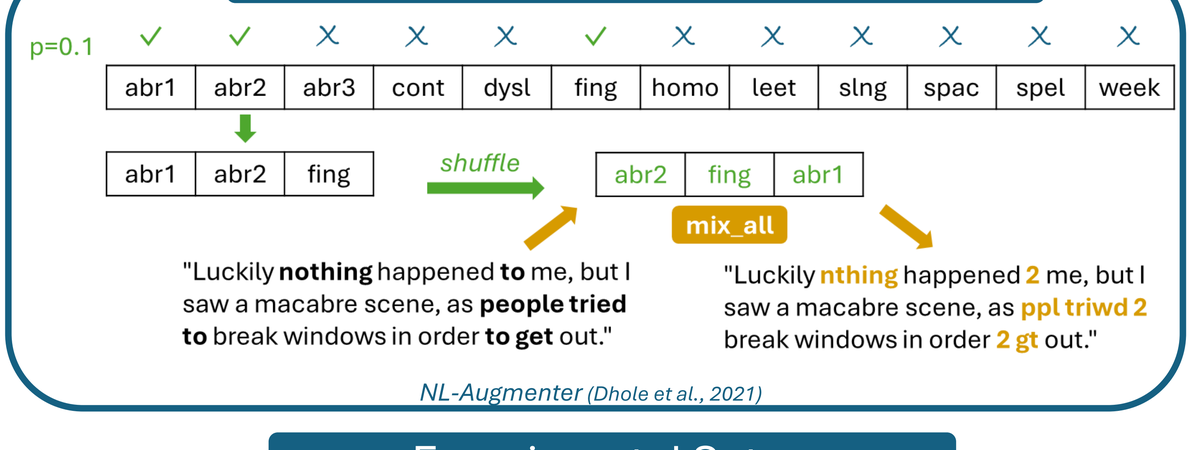
Teacher-Student Approach



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

Full paper here

Generating Artificial UGC Training Data



NL-Augmenter (Dhole et al., 2021)

Experimental Setup

Models

- **LASER (teacher)**
  - 5-layer bi-LSTM (45M params)
  - fixed during training
- **RoLASER [Robust LASER] (student)**
  - 12-layer Transformer (108M params)
  - initialised with RoBERTa *(Liu et al., 2019)*
- **c-RoLASER (student)**
  - 12-layer Transformer (104M params)
  - initialised with CharacterBERT *(El Boukkouri et al., 2020)*

Metrics

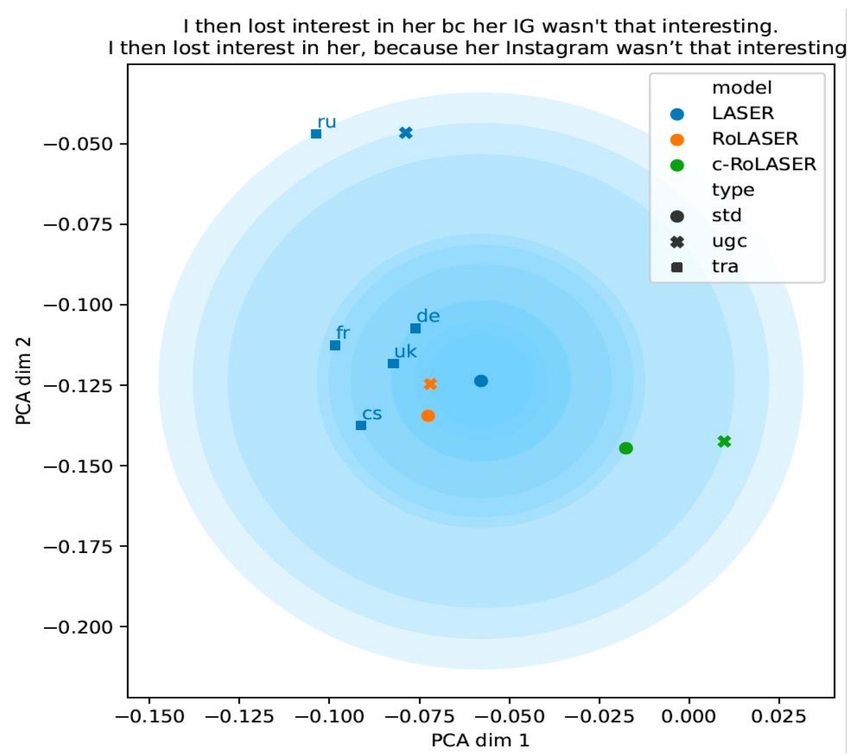
- **Average pairwise cosine distance**
- **xSIM** *(Artetxe and Schwenk, 2019)*
  - proxy metric for bitext mining
  - cross-lingual similarity search
  - error rate of aligning translation pairs
- **xSIM++** *(Chen et al., 2023)*
  - more challenging than xSIM

Data

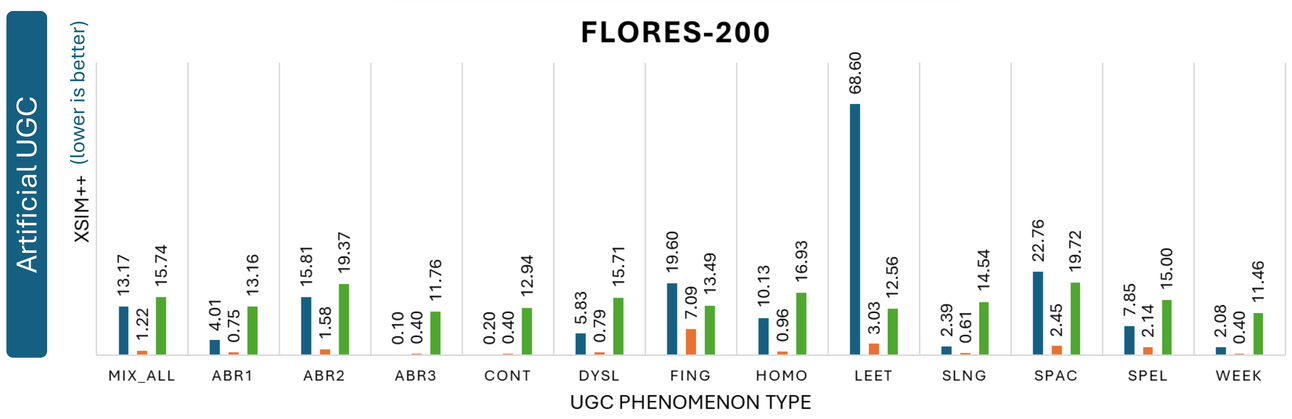
- **OSCAR** *(Ortiz Suárez et al., 2019)*
  - 2M lines of standard English used for training
  - artificially augmented with the *mix_all* transformation
- **RoCS-MT** *(Bawden and Sagot, 2023)*
  - 1922 standard ↔ UGC English sentences from Reddit
  - translations into 5 other languages
- **FLORES-200** *(NLLB Team et al., 2022)*
  - 1012 standard English sentences from WikiNews, WikiBooks, WikiVoyage
  - artificially augmented with *mix_all* and other NL-Augmenter transformations

Results and Analysis

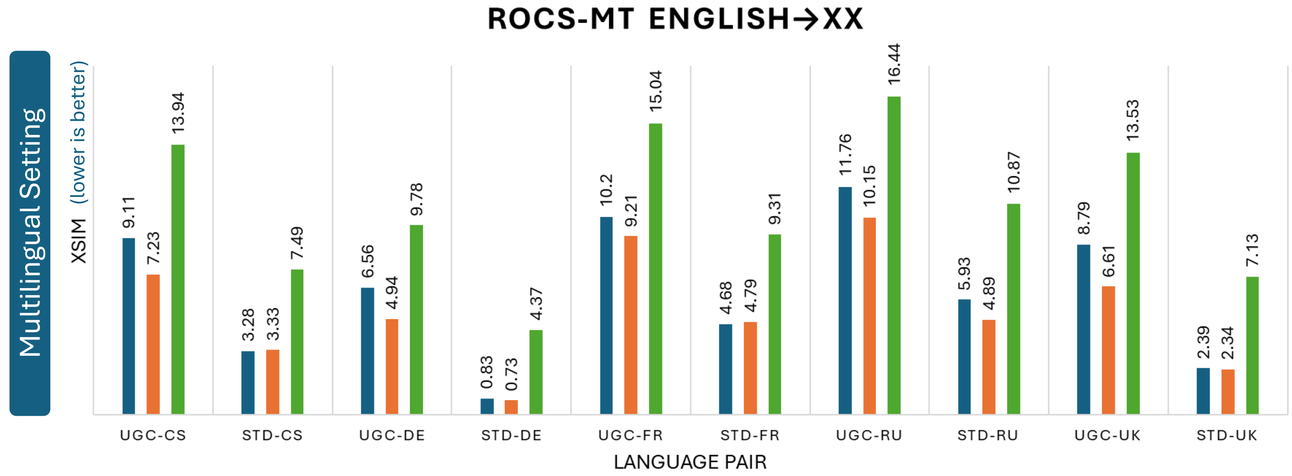
Natural UGC: RoCS-MT



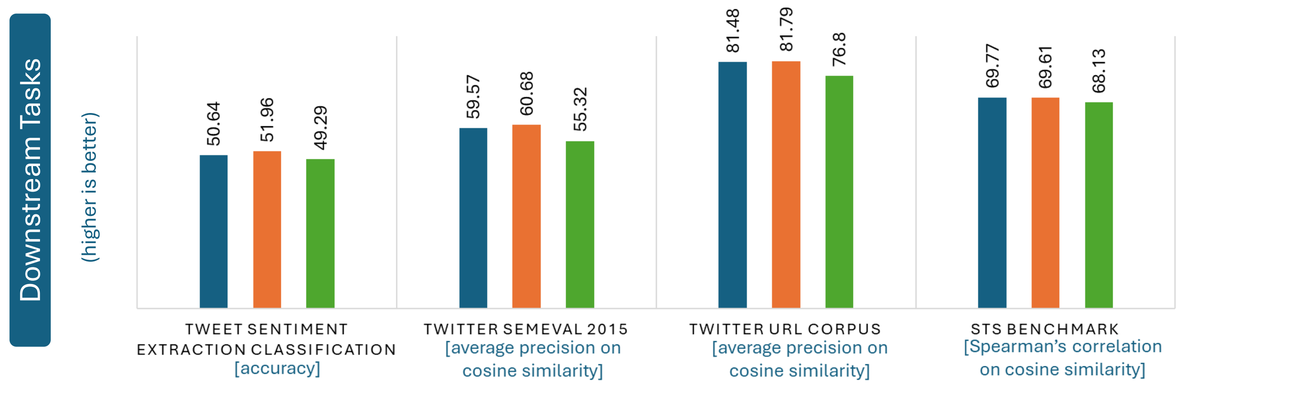
Artificial UGC



Multilingual Setting



Downstream Tasks



RoLASER outperforms LASER on UGC

c-RoLASER struggles to map its standard embeddings to LASER's

The **most challenging UGC** phenomena **shatter subword tokenisation**

Future work:
Extend RoLASER to **more languages**