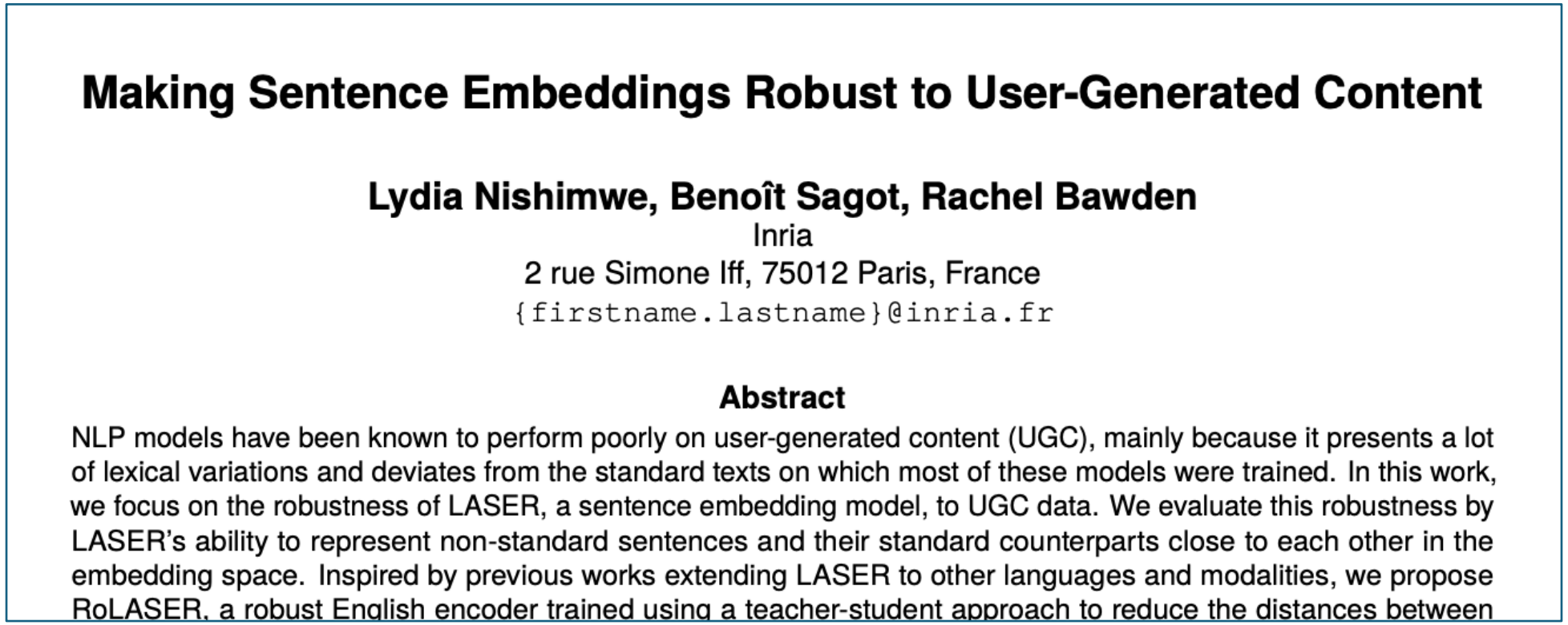
Making Sentence Embeddings Robust to User-Generated Content

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

MARI Seminar - May 29, 2024

LREC-COLING 2024

Lydia Nishimwe, Benoît Sagot, and Rachel Bawden. 2024. **Making Sentence Embeddings Robust to User-Generated Content.** In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10984–10998, Torino, Italia. ELRA and ICCL.



Making Sentence Embeddings Robust to User-Generated Content

Lydia Nishimwe, Benoît Sagot, Rachel Bawden
Inria
2 rue Simone Iff, 75012 Paris, France
{firstname.lastname}@inria.fr

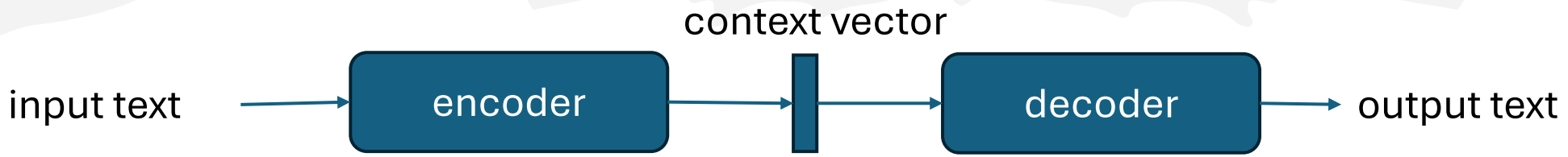
Abstract

NLP models have been known to perform poorly on user-generated content (UGC), mainly because it presents a lot of lexical variations and deviates from the standard texts on which most of these models were trained. In this work, we focus on the robustness of LASER, a sentence embedding model, to UGC data. We evaluate this robustness by LASER's ability to represent non-standard sentences and their standard counterparts close to each other in the embedding space. Inspired by previous works extending LASER to other languages and modalities, we propose RoLASER, a robust English encoder trained using a teacher-student approach to reduce the distances between

I. Introduction

Background and Motivation

Natural Language Processing (NLP)



Encoder-Decoder Tasks

- Machine translation
- Text summarisation
- Question answering

e.g. Bing Translator

Encoder-only Tasks

- Text classification
- Named Entity Recognition (NER)
- Part-of-Speech (PoS) Tagging
- Textual Entailment

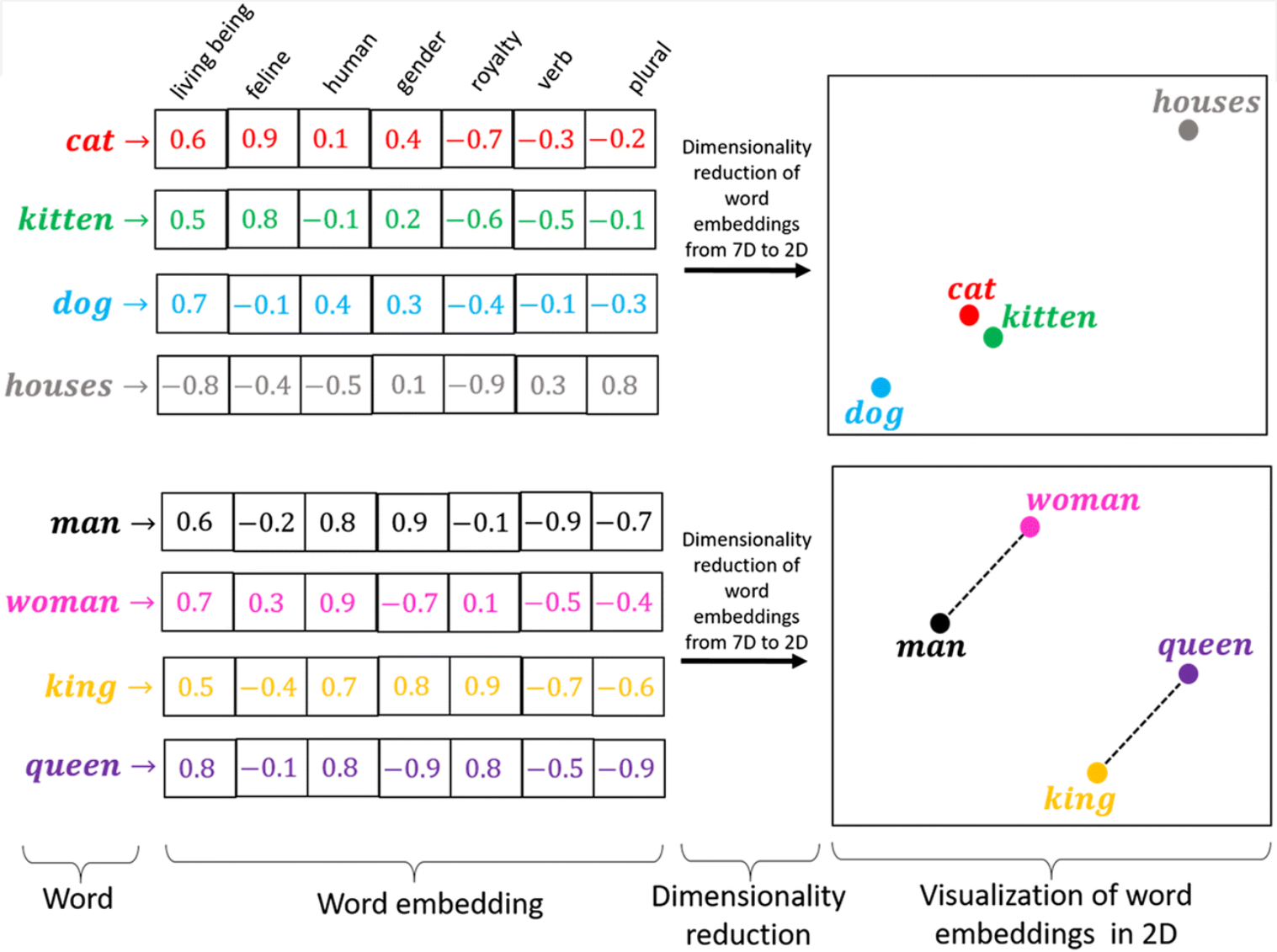
e.g. BERT

Decoder-only Tasks

- Text generation/completion
- Language modelling
- Code generation

e.g. GPT

Word embeddings



(Hariom Gautam, 2020)

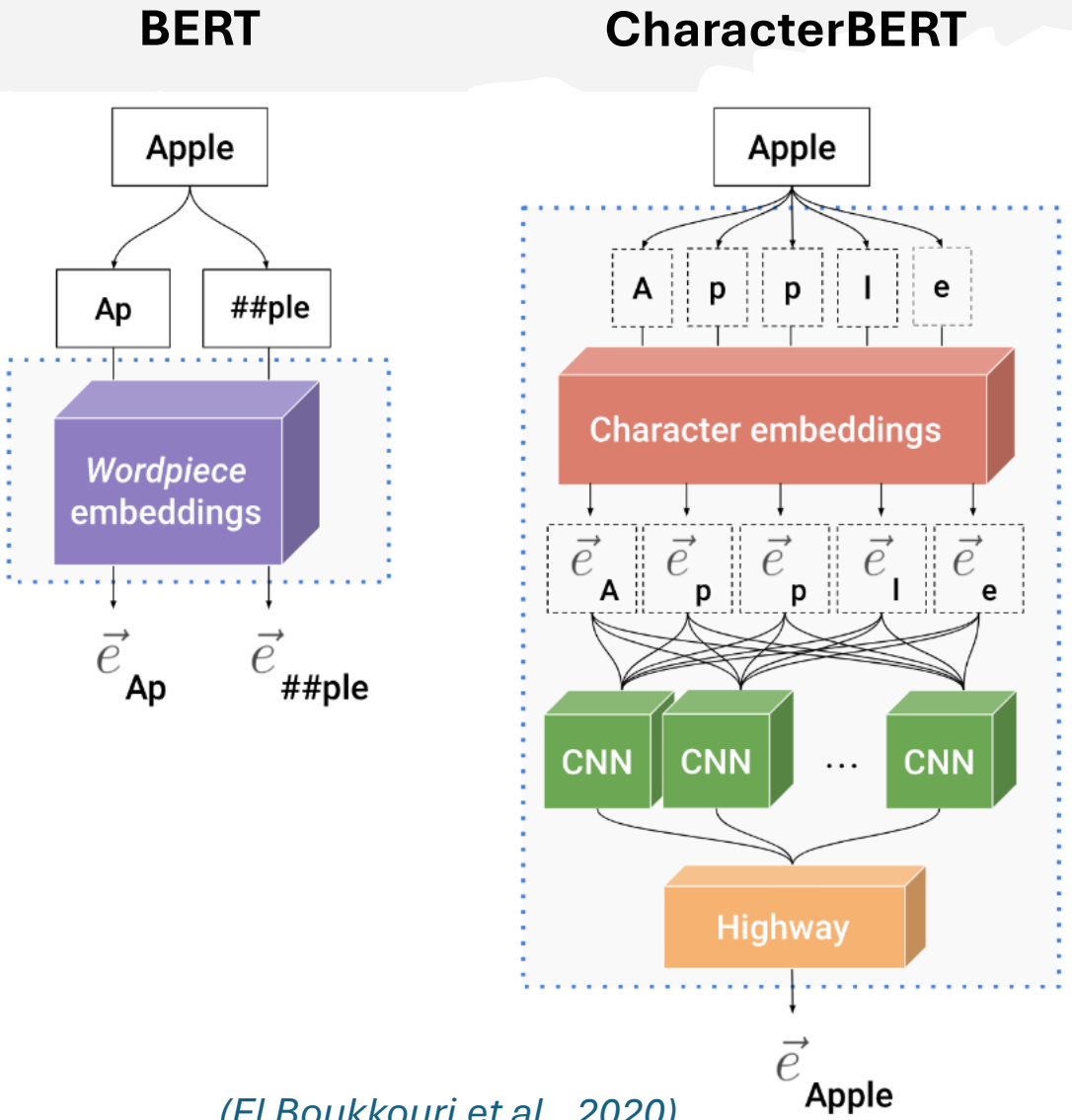
Tokenisation

This is a sentence.

words: This is a sentence .

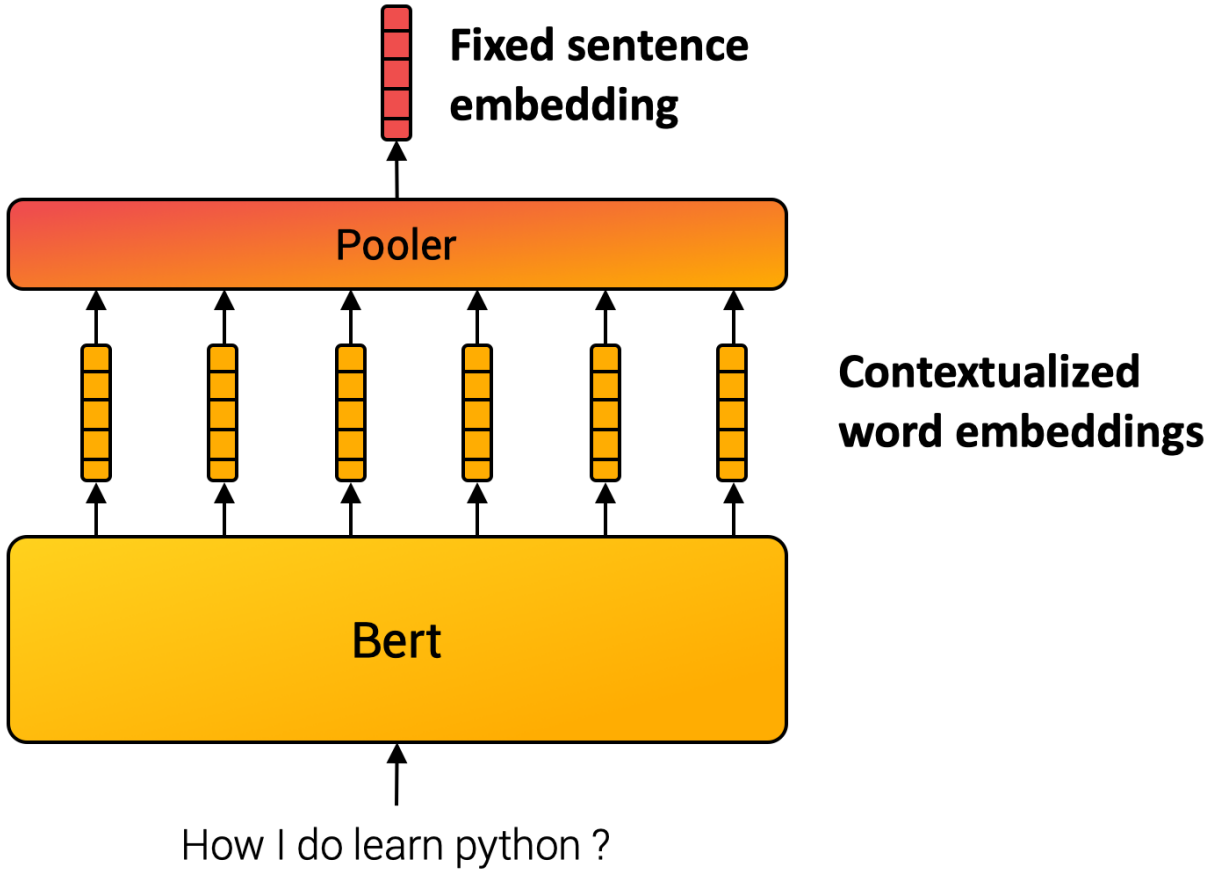
subwords: This is a sent ##ence .

characters: T h i s _ i s _ a _ s e n t e n c e _ .



(El Boukkouri et al., 2020)

Sentence embeddings

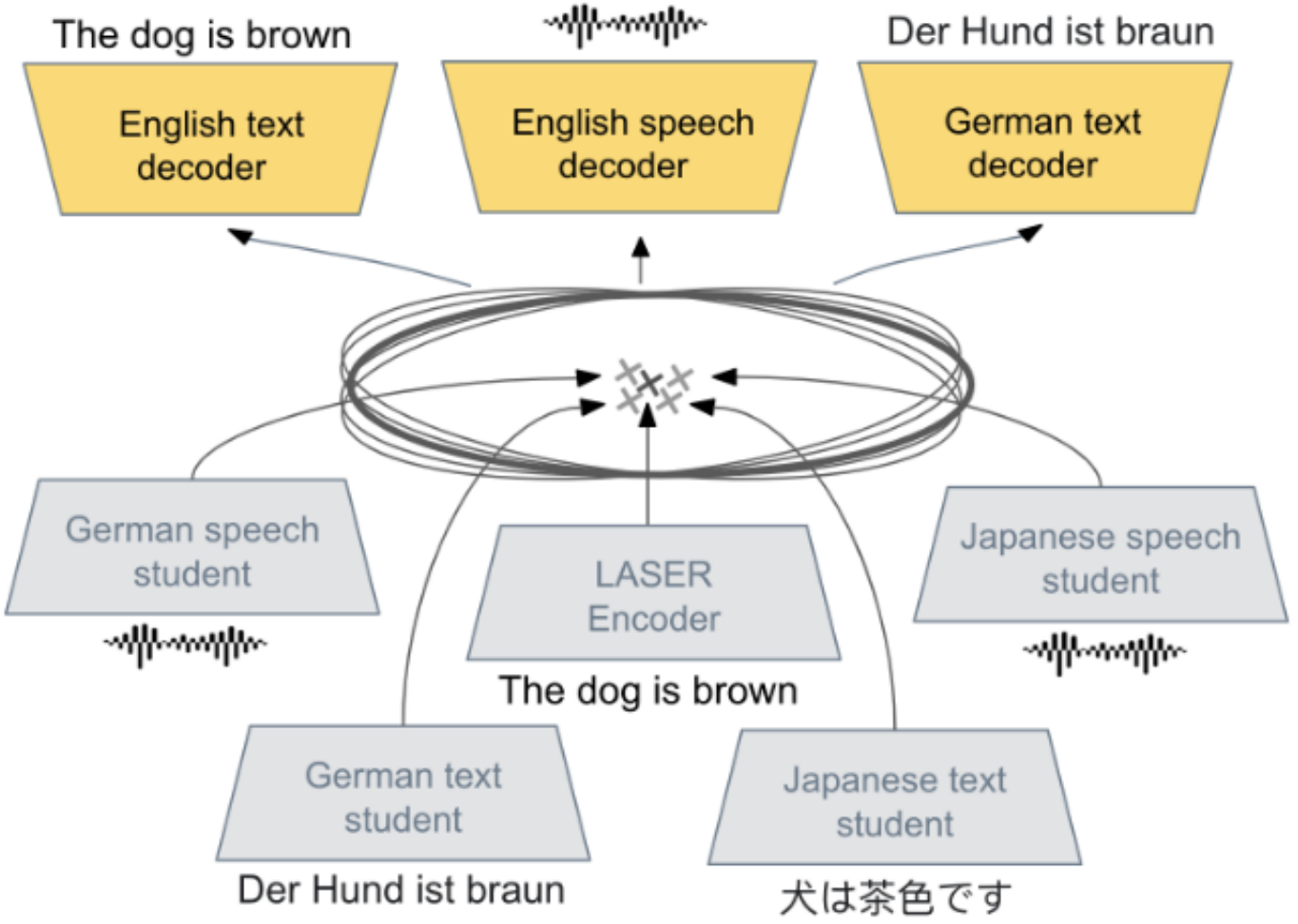


(Antoine Simoulin, 2021)

Applications

- Semantic Textual Similarity (STS)
  - Plagiarism detection
  - Document clustering
- Bitext Mining
- Text Classification
  - Sentiment analysis
  - Spam detection
  - Topic classification
- Text Pair Classification
  - Paraphrase Identification
- Information Retrieval (IR)
  - Search engines
  - Question answering

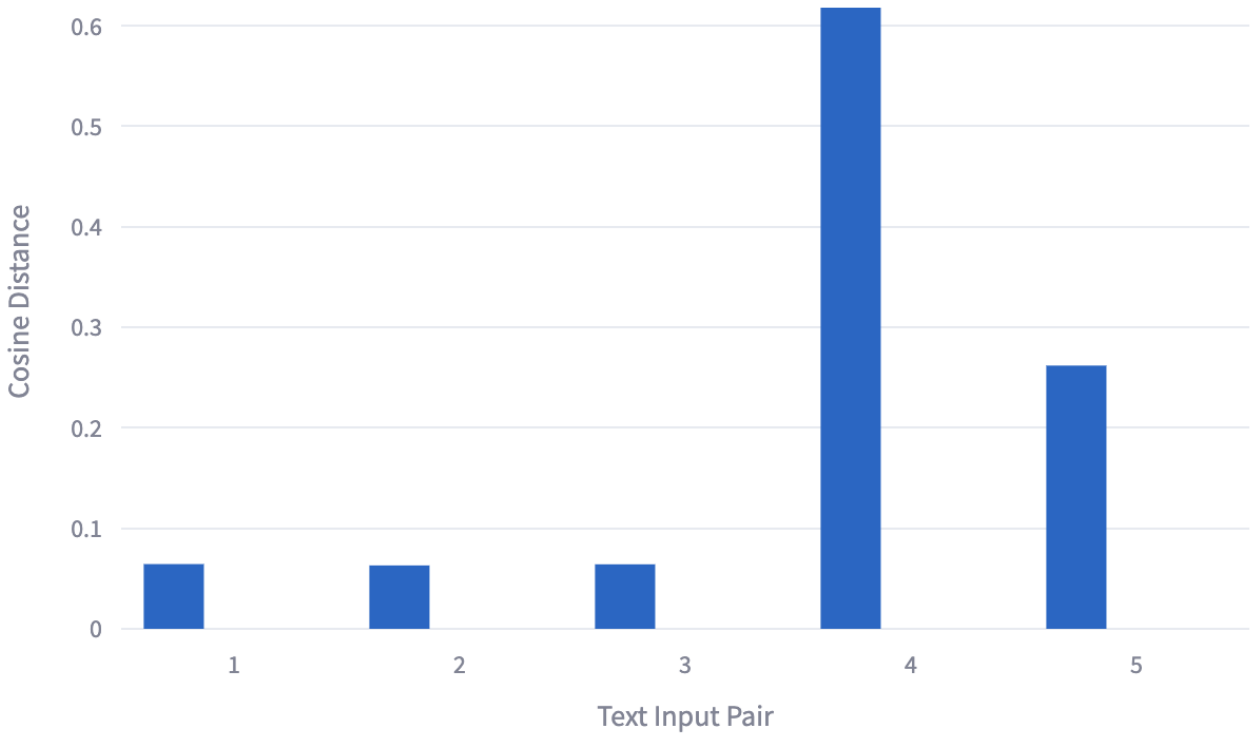
LASER: **L**anguage-**A**gnostic **SE**ntence **R**epresentations



(Artetxe and Schwenk, 2019)
(Heffernan et al., 2022)
(Duquenne et al., 2022)

LASER’s multilingual embeddings

Standard text	Non-standard text
Standard text 1: See you tomorrow.	Non-standard text 1: À demain.
Standard text 2: See you tomorrow.	Non-standard text 2: Bis morgen.
Standard text 3: See you tomorrow.	Non-standard text 3: Hasta mañana.
Standard text 4: See you tomorrow.	Non-standard text 4: Tuonane kesho.
Standard text 5: See you tomorrow.	Non-standard text 5: Let's meet up tomorrow.



RoLASER Demo App

User-Generated Content (UGC)

Ergographic phenomena (encoding simplification)

i **don wanna fyt witchu**

al b an **our l8**

c u 2moro

Neologisms

The math is not **mathing**.

burkini

Transverse phenomena

i **aint playin**

idk

afaik

N. E. V. E. R

Foreign language influence

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

likez et commentez

Marks of expressiveness

super**rrrr !!!!**

<3

☺

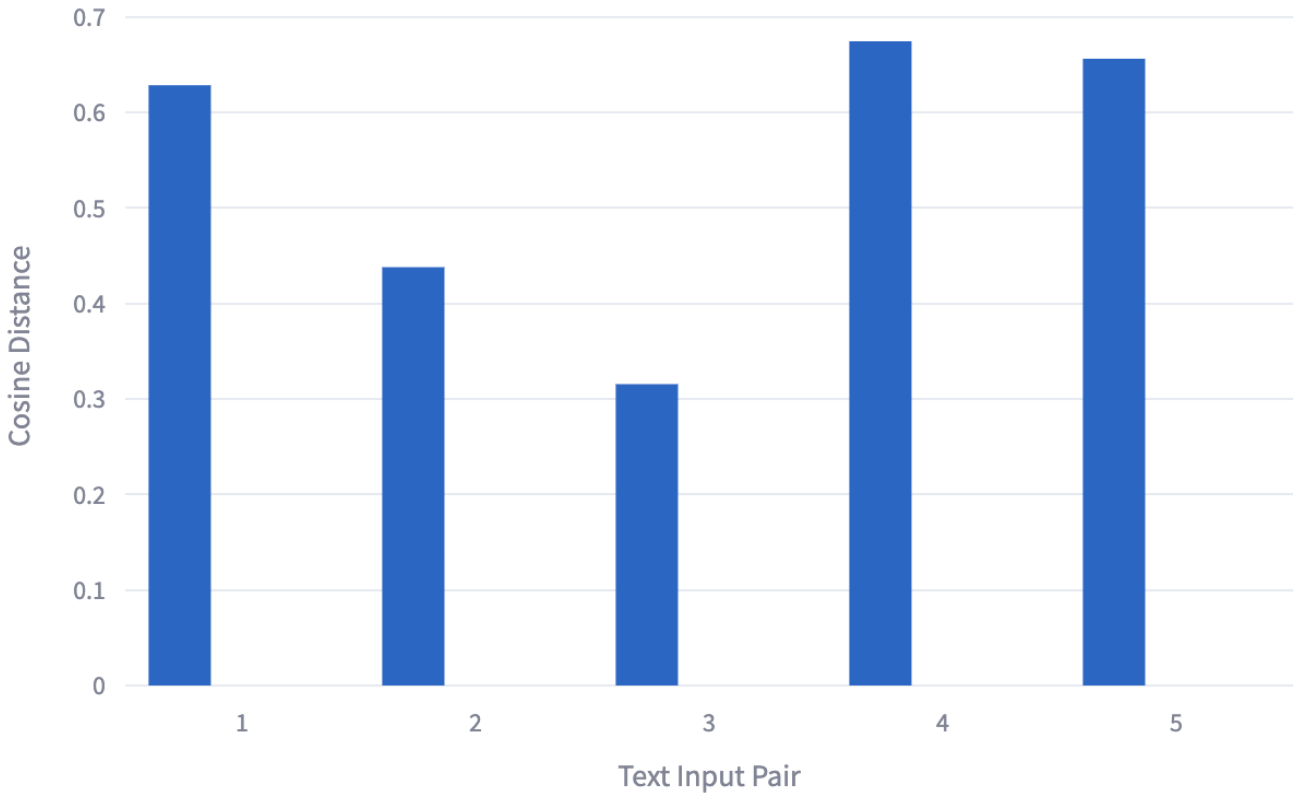
!d10t

sh*t

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

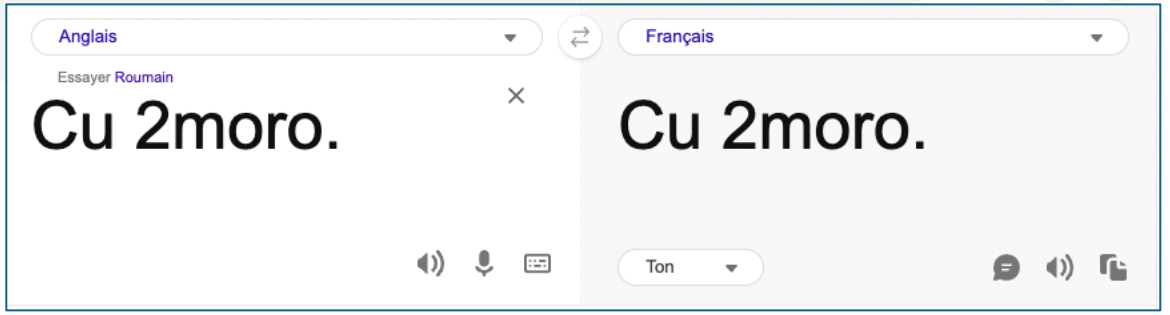
LASER’s UGC embeddings

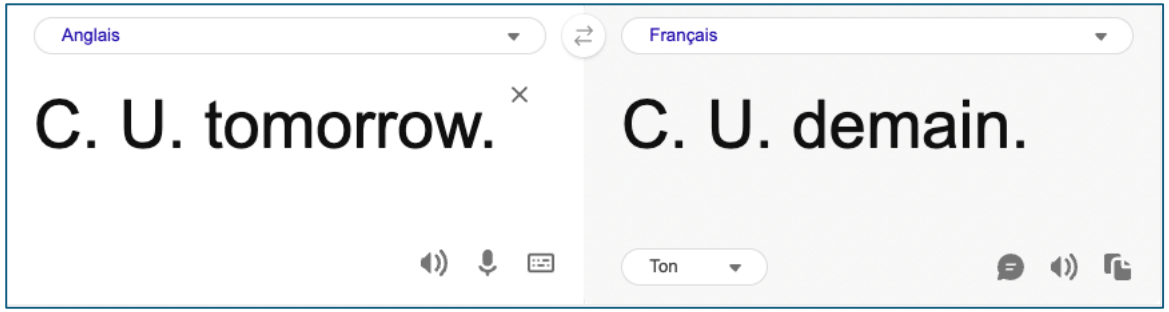
Standard text 1:	Non-standard text 1:
See you tomorrow.	See you t03orro3.
Standard text 2:	Non-standard text 2:
See you tomorrow.	C. U. tomorrow.
Standard text 3:	Non-standard text 3:
See you tomorrow.	sea you tomorrow.
Standard text 4:	Non-standard text 4:
See you tomorrow.	See yo utomorrow.
Standard text 5:	Non-standard text 5:
See you tomorrow.	Cu 2moro.

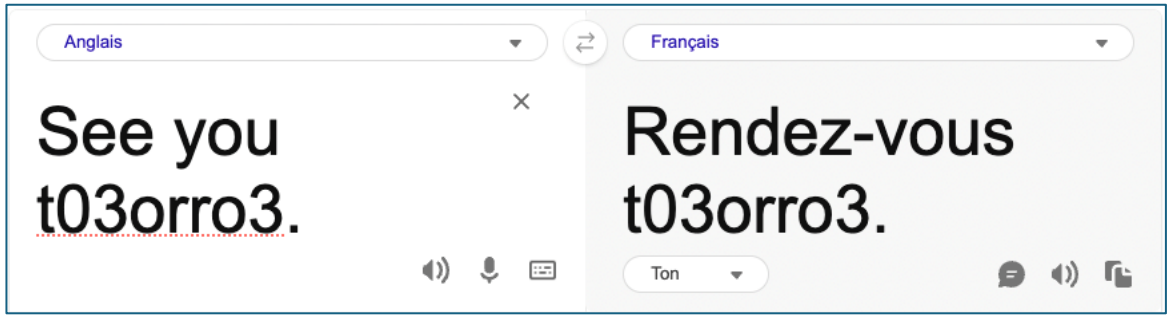


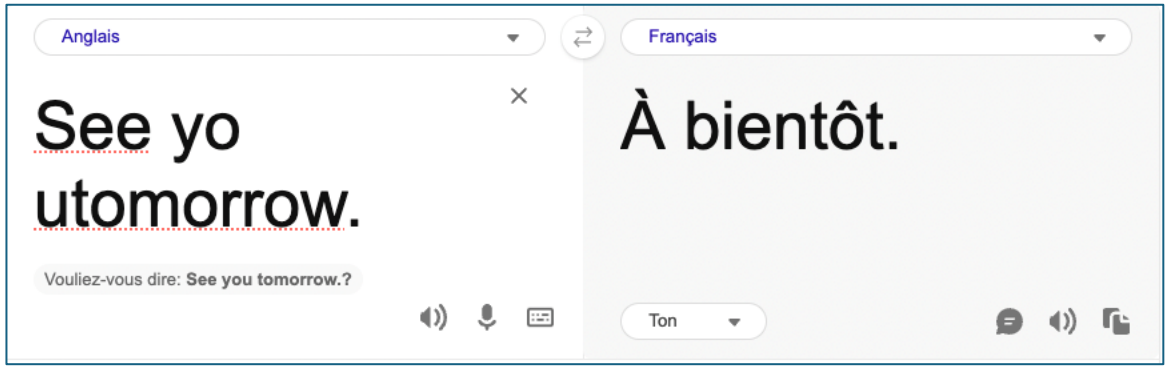
RoLASER Demo App

Negative effects of UGC



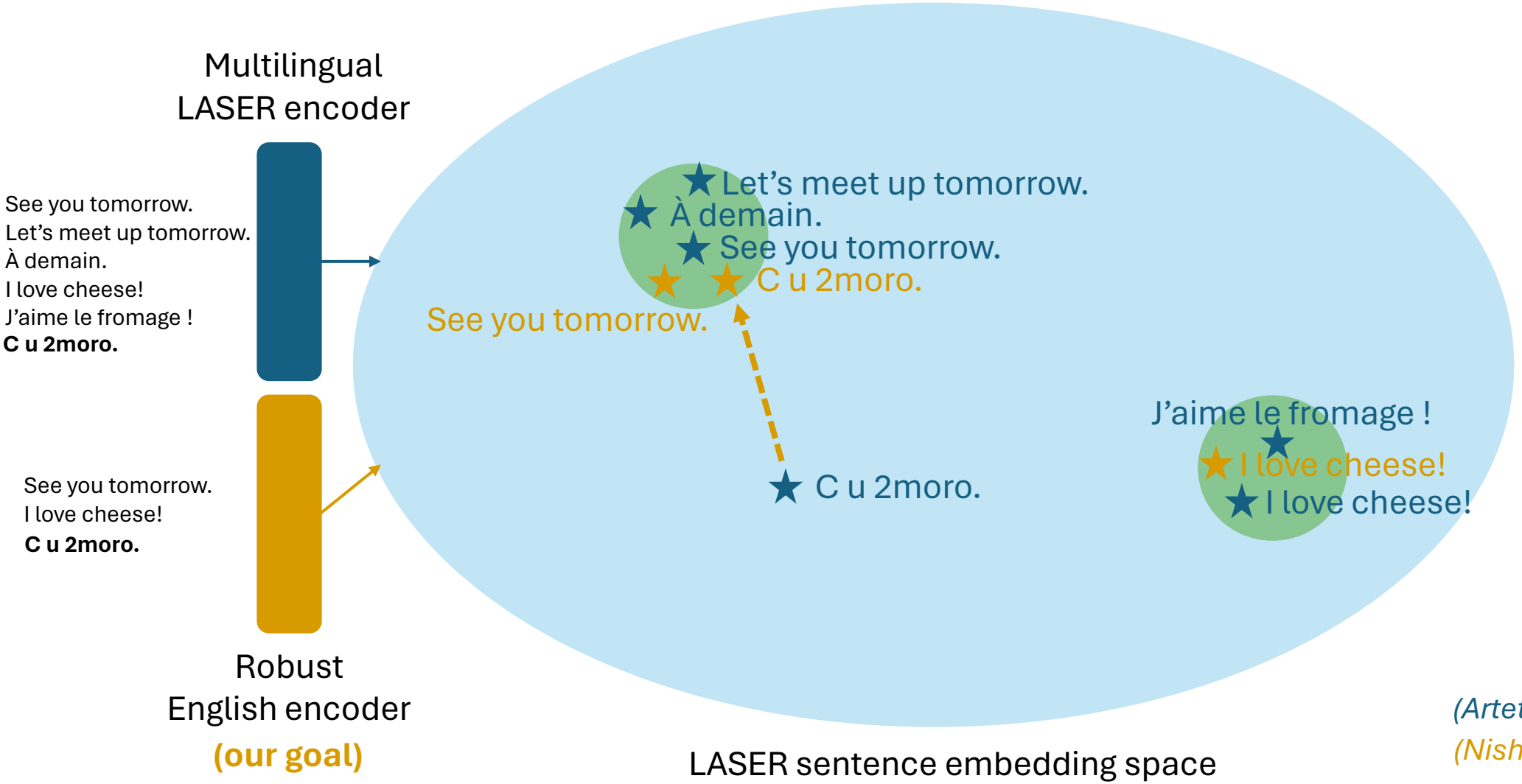






Microsoft Bing Translator

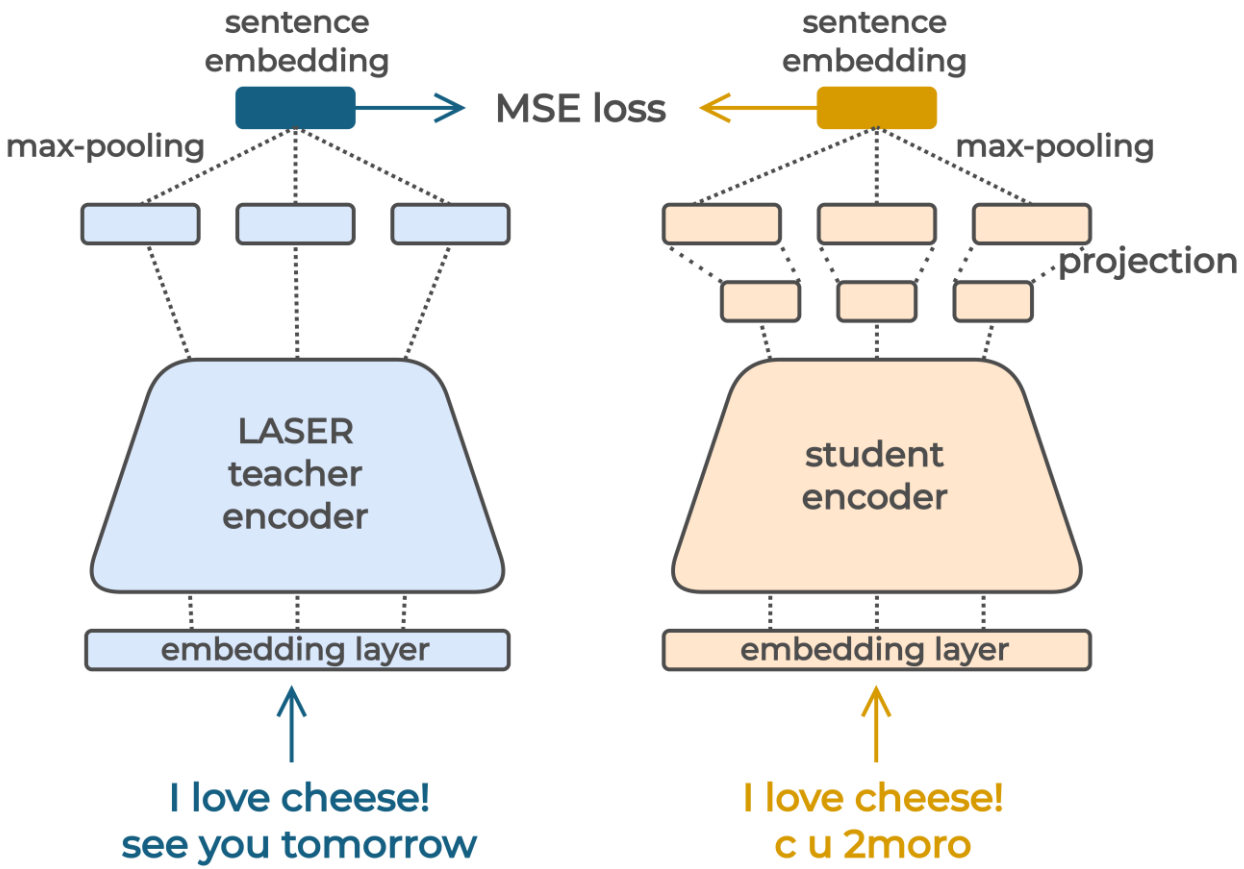
Multilingual sentence embeddings



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

II. 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

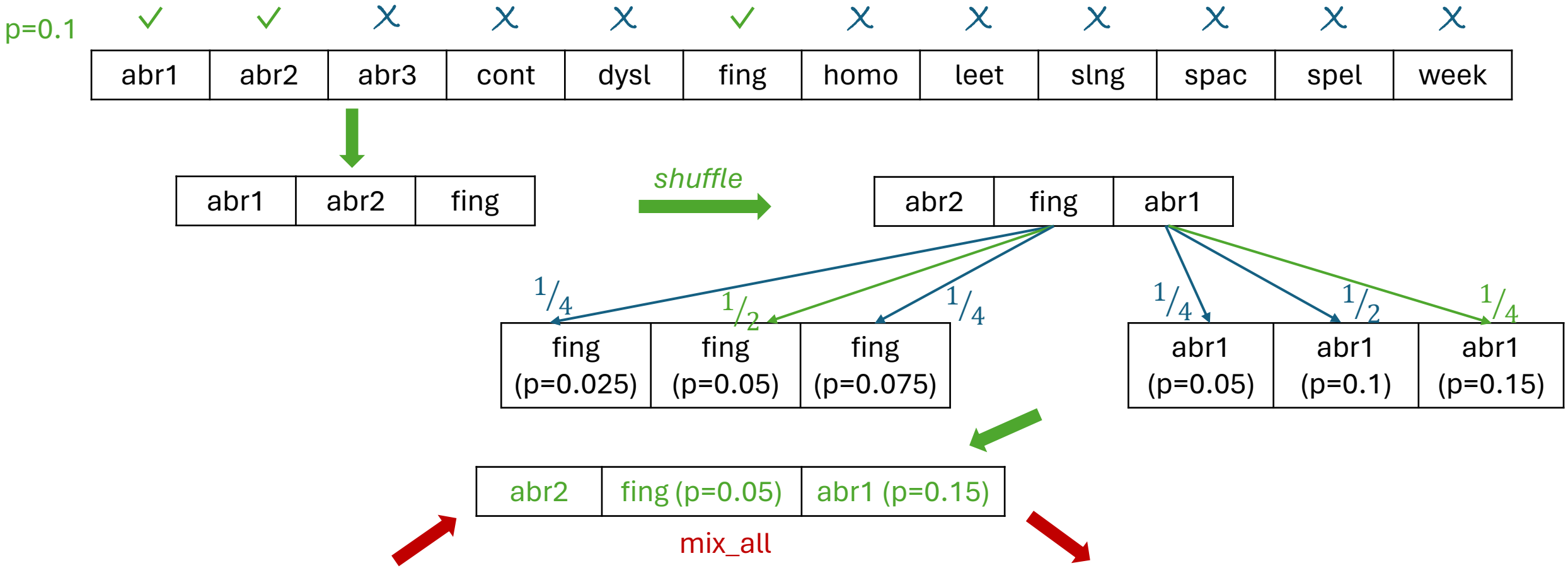
homo there ↔ their

dysl lose ↔ loose

spel absent → apsent

(Dhole et al., 2021)

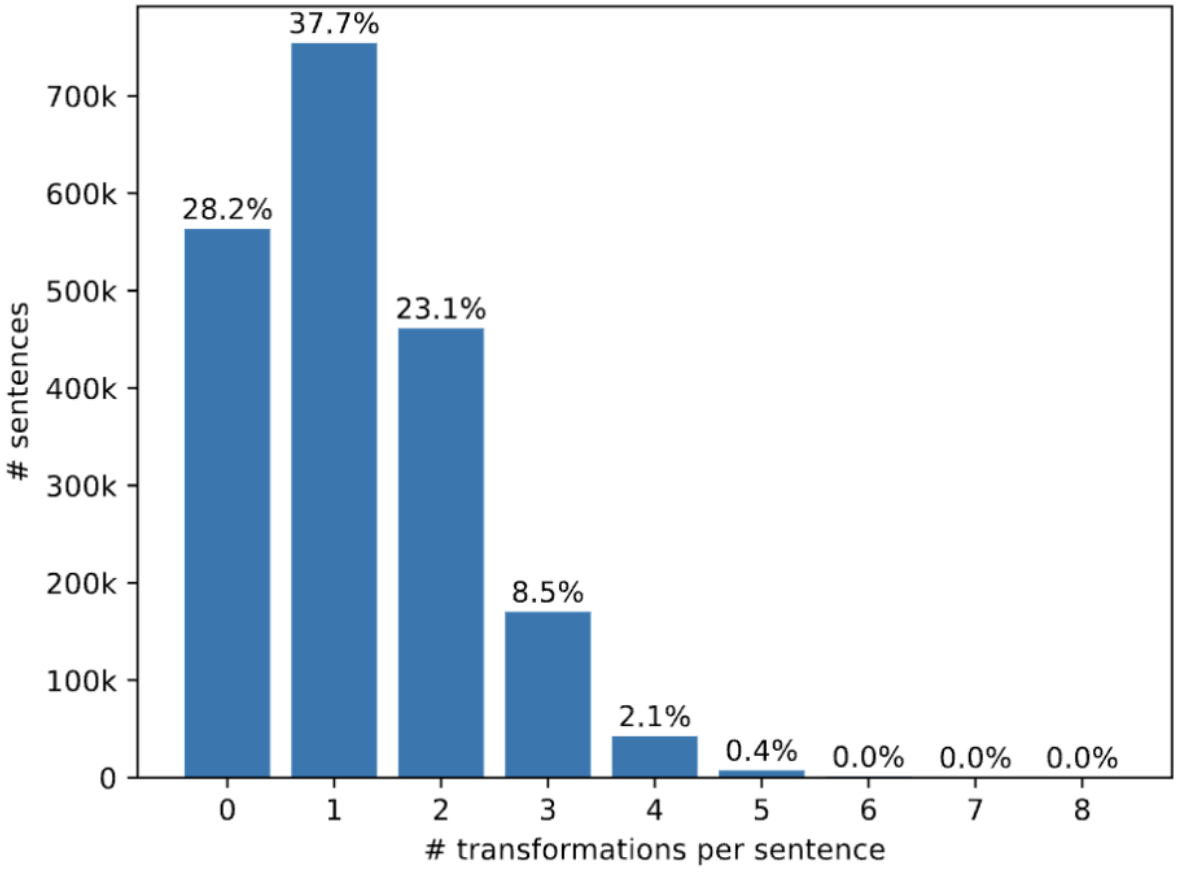
Generating artificial UGC training data

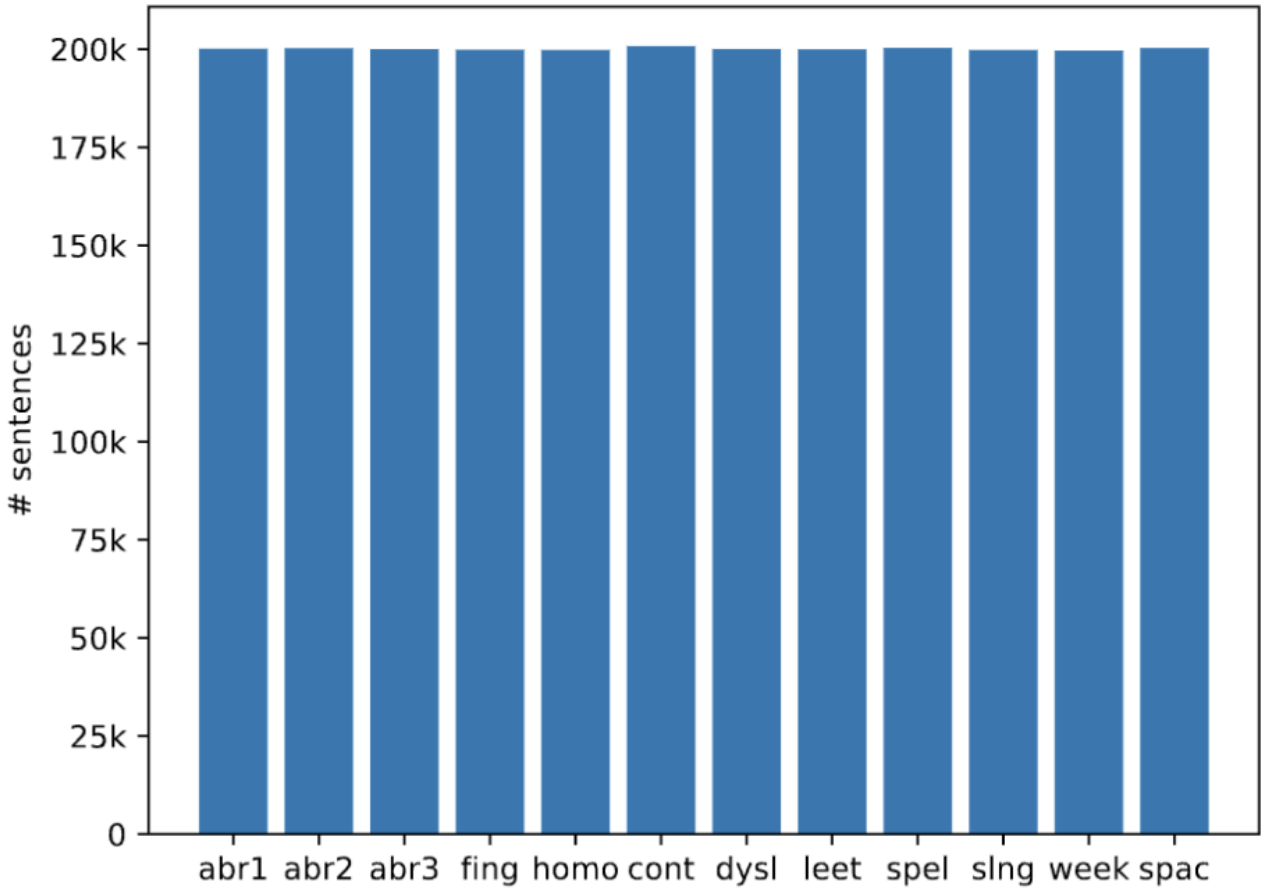


"Luckily **nothing** happened **to** me, but I saw a macabre scene, as **people tried to** break windows in order **to get** out."

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

Artificial UGC training data





III. Experiments

Evaluation data

Corpus	UGC sentence	Standard(ised) sentence
MultiLexNorm$^{\diamond}$	if i **cnt** afford the real deal , i ain't buying **nuffin** fake .. i just won't have it	if i **can't** afford the real deal , i ain't buying **nothing** fake .. i just won't have it
RoCS-MT‡	**Umm idk**, maybe **its bc** we're **DIFFERENT PEOPLE** with **DIFFERENT BODIES???**	**Um, I don't know**, maybe **it's because** we're **different people** with **different bodies?**
FLORES† `abr2 + fing + abr1`	**"** Luckily **nthing** happened **2** me **,** but I saw a macabre scene **,** as **ppl triwd 2** break windows in order **2 gt** out **.**	**"**Luckily **nothing** happened **to** me**,** but I saw a macabre scene**,** as **people tried to** break windows in order **to get** out**.**

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

Experimental setup

- **Training data:**
  - 2M “bilingual” standard-UGC lines
  - 2M standard English lines from the OSCAR dataset *(Ortiz Suárez et al., 2019)*
  - augmented with the *mix_all* transformation
- **Validation data:**
  - FLORES-200 dev set + *mix_all*

- **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 metrics

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

How closely the models embed non-standard sentences to their standard counterparts

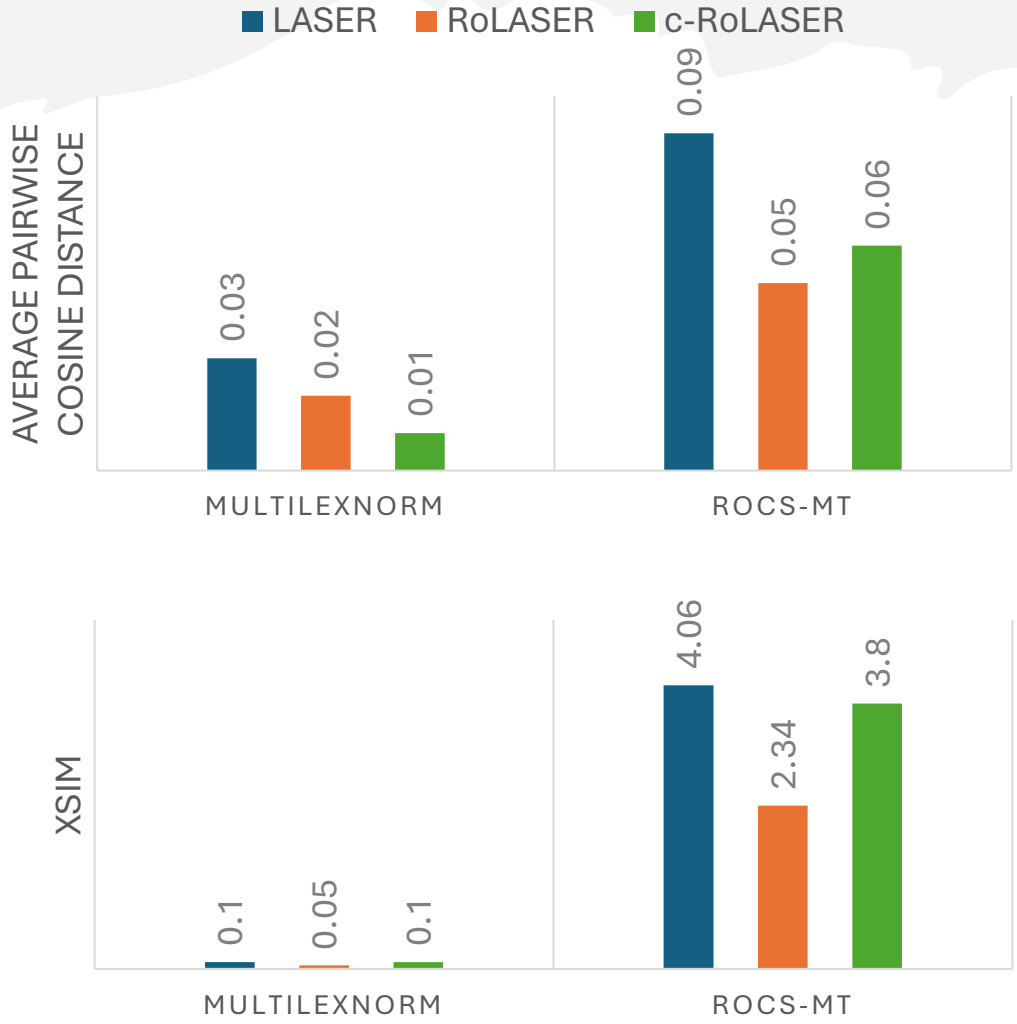
How well the models align non-standard sentences to their standard counterparts

Evaluating robustness

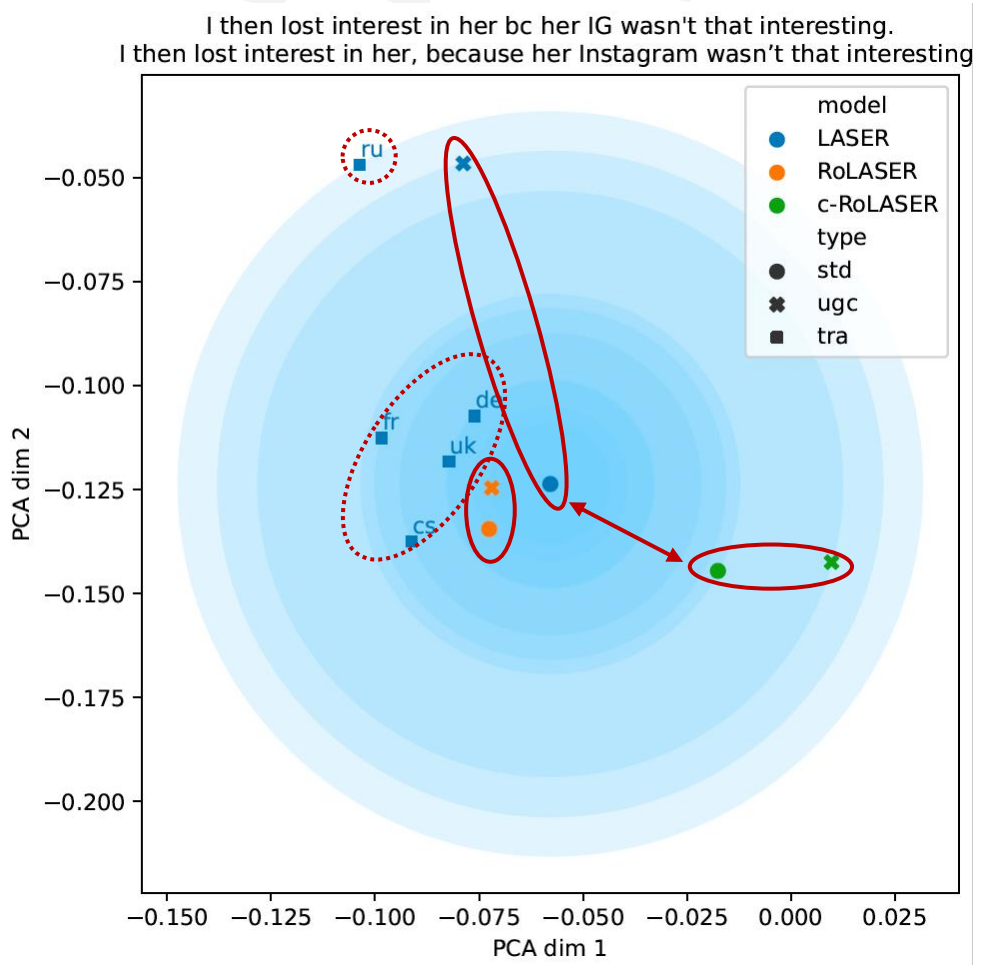
1. Does robustness to artificial UGC translate to robustness to natural UGC?
2. Can the students replace LASER at representing English sentences in a multilingual setting?
3. Does robustness to UGC degrade performance on standard data?
4. Does robustness in sentence embeddings impact performance on downstream tasks?

IV. Results and Analysis

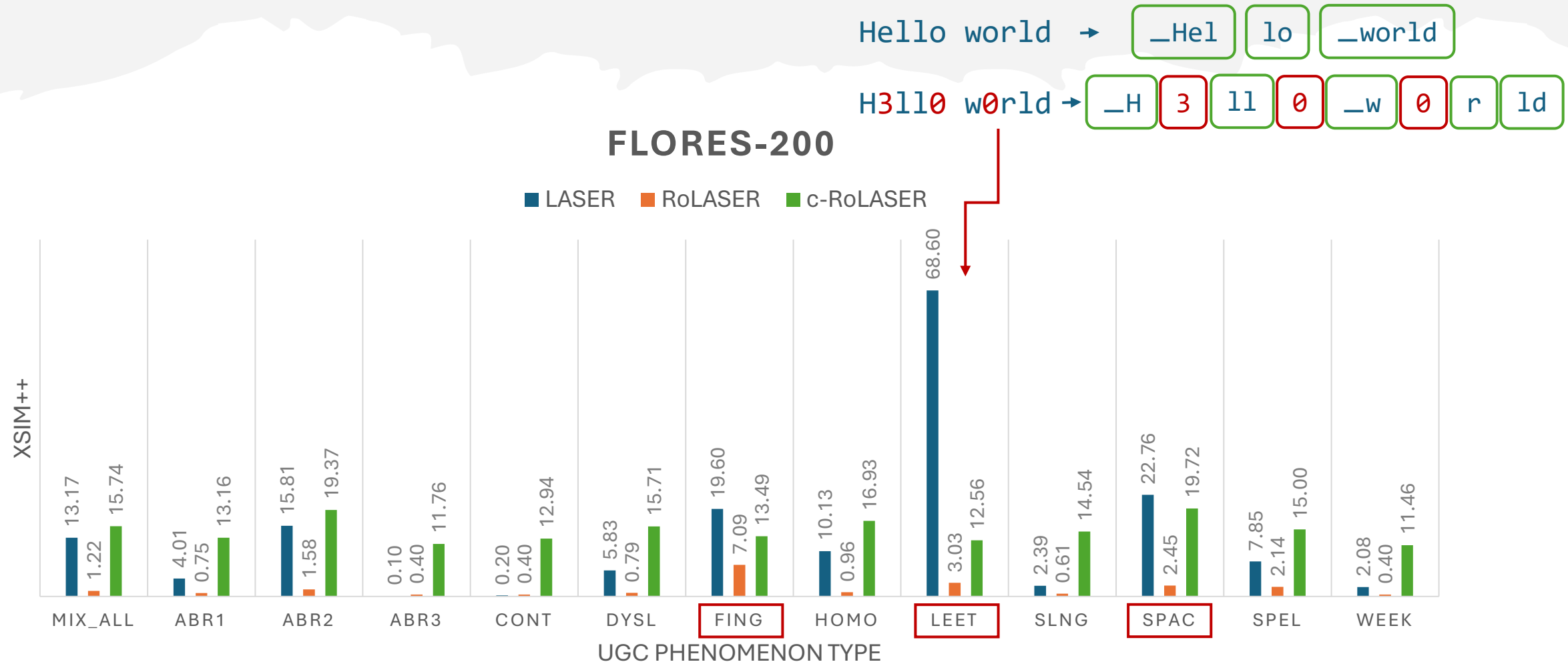
Evaluation on natural UGC



(lower is better)

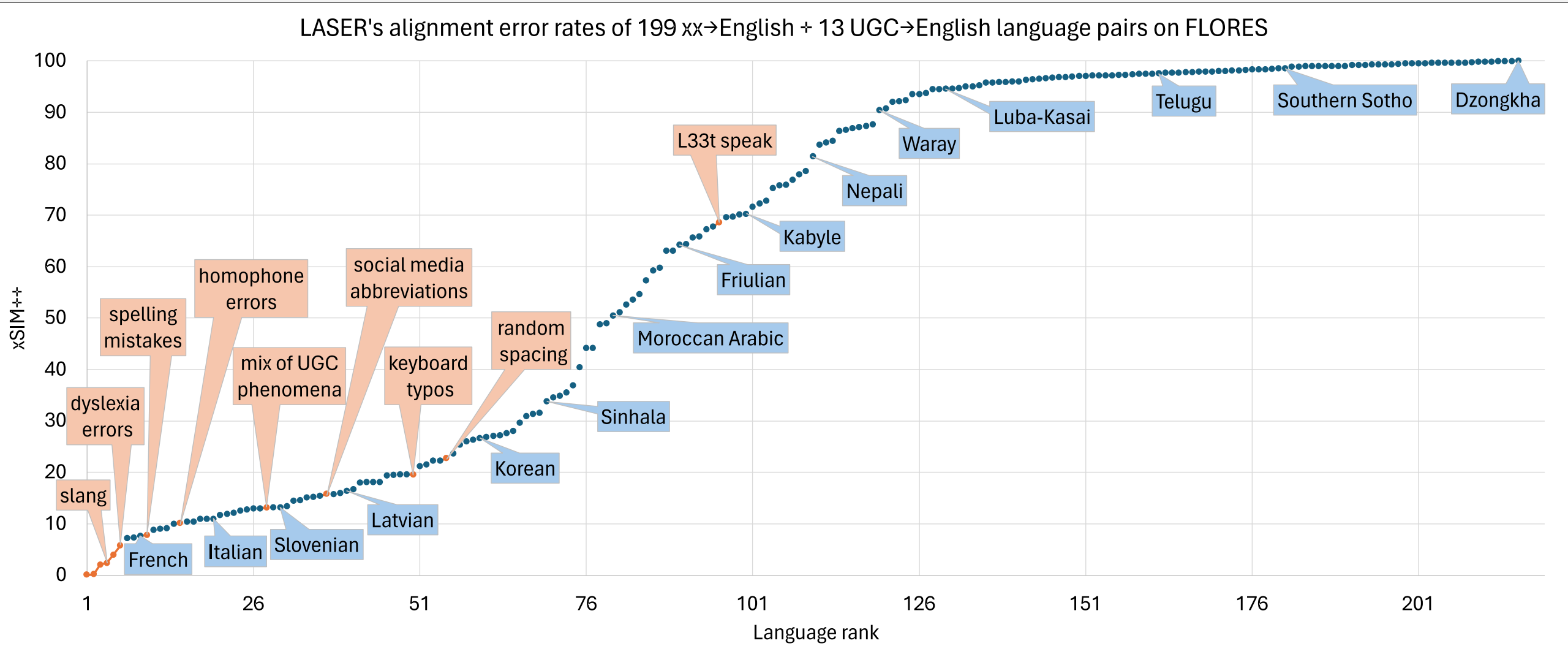


Evaluation on artificial UGC



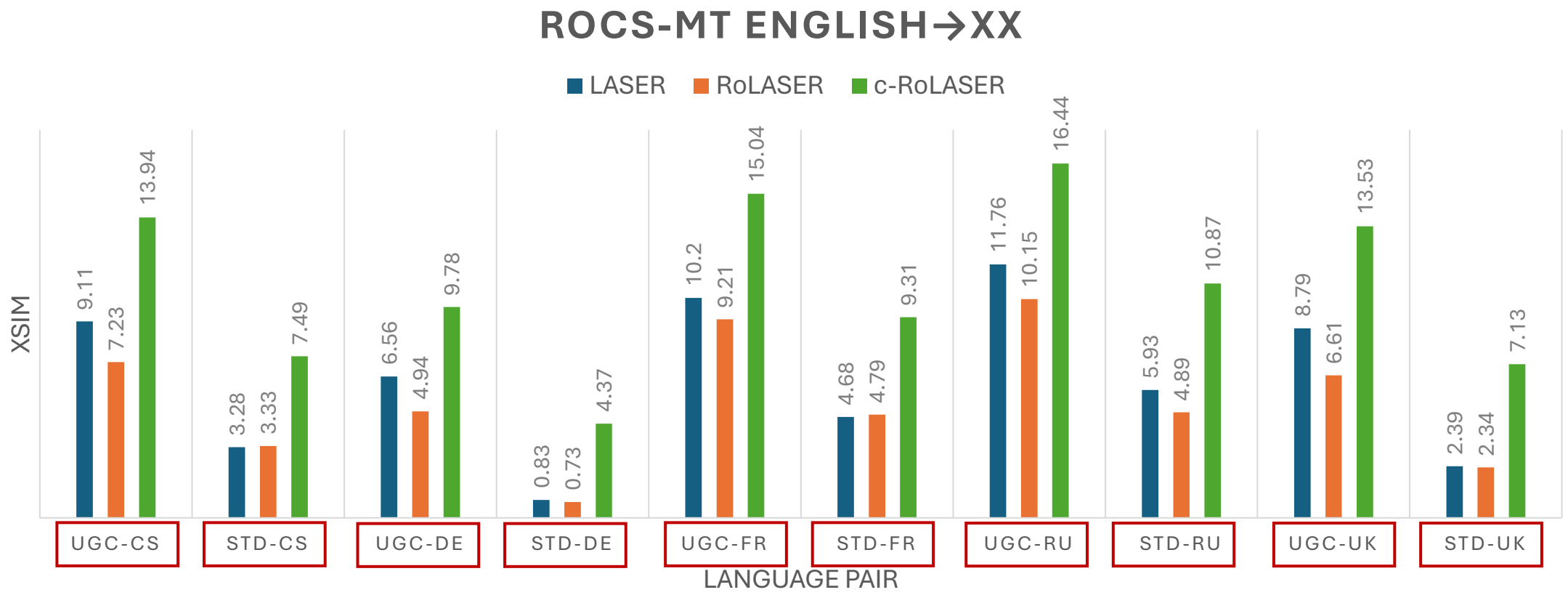
(lower is better)

LASER's embeddings of UGC and other languages



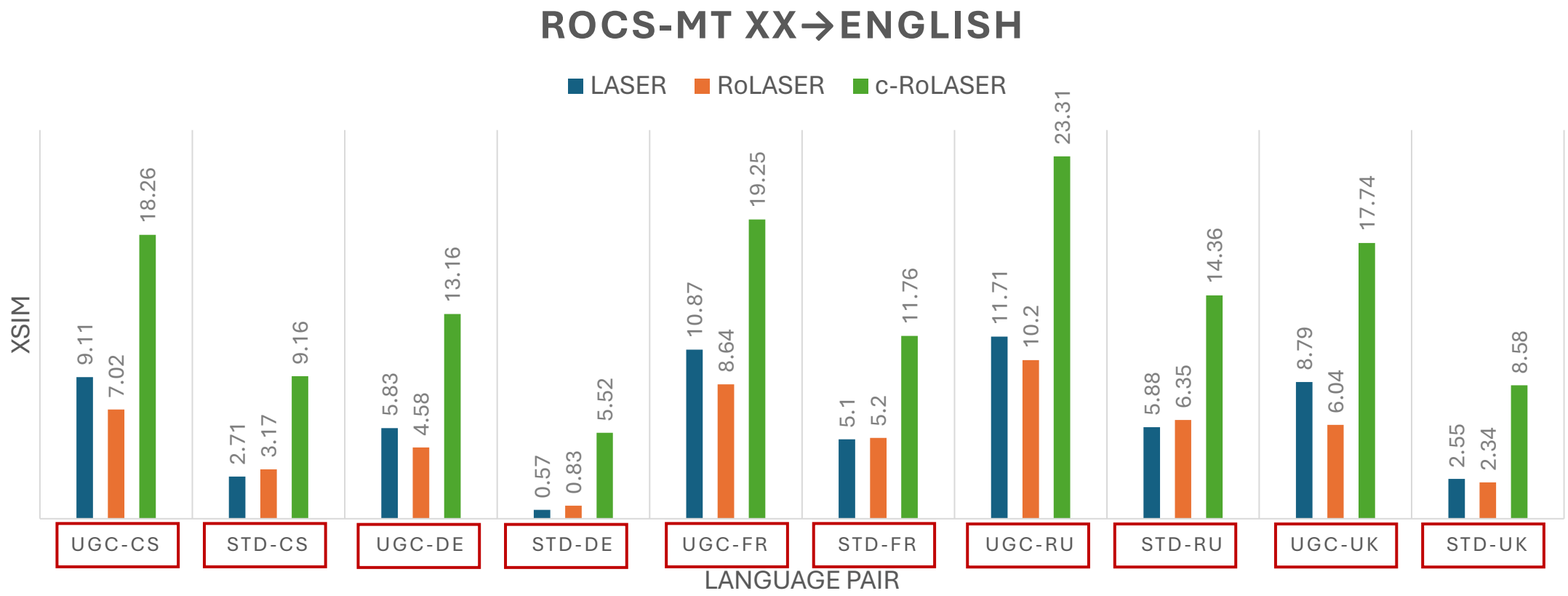
(Lower is better)

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)

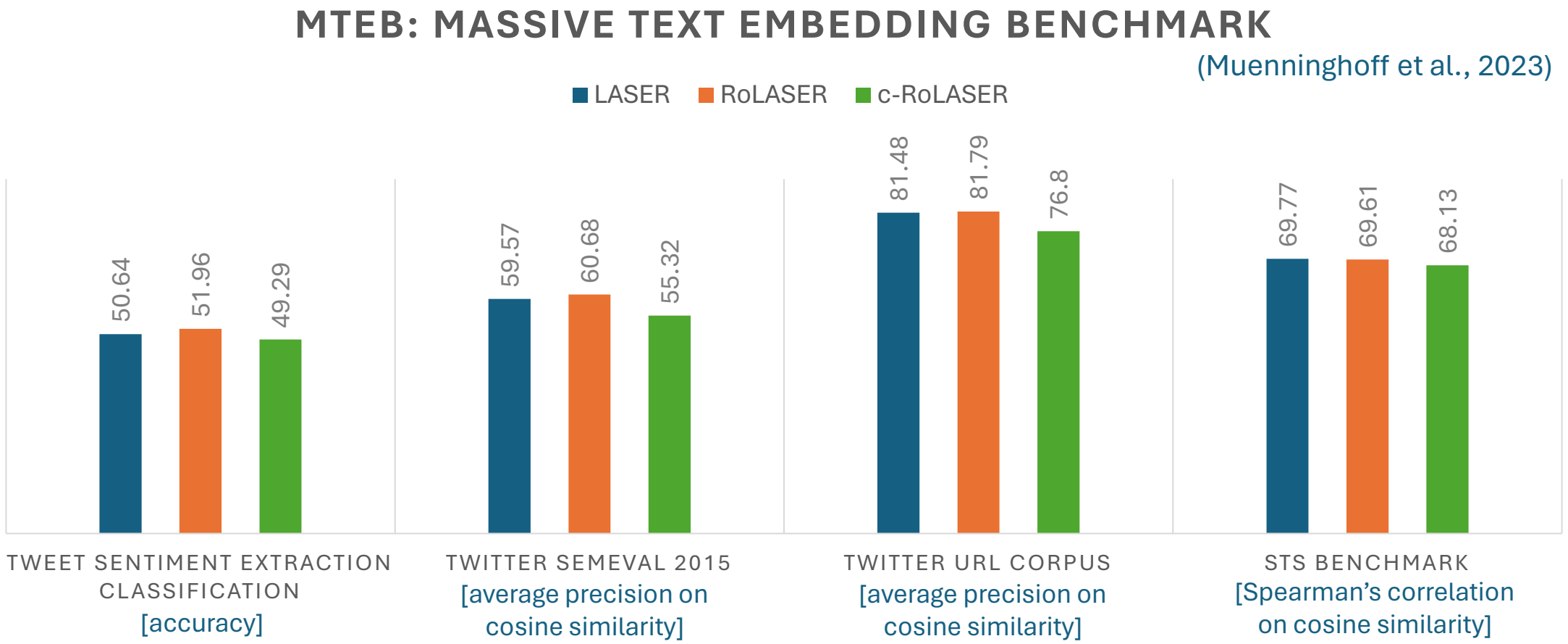


(lower is better)

Evaluation on downstream tasks (1)

1. **Sentence classification**, which predicts labels from sentence embeddings, e.g. sentiment labels:
   - Tweet Sentiment Extraction Classification
2. **Sentence pair classification**, which predicts a binary label from sentence embeddings, e.g. whether two sentences are paraphrases:
   - Twitter Sem Eval 2015
   - Twitter URL Corpus
3. **Semantic textual similarity**, which examines the degree of semantic equivalence between two sentences:
   - STS Benchmark

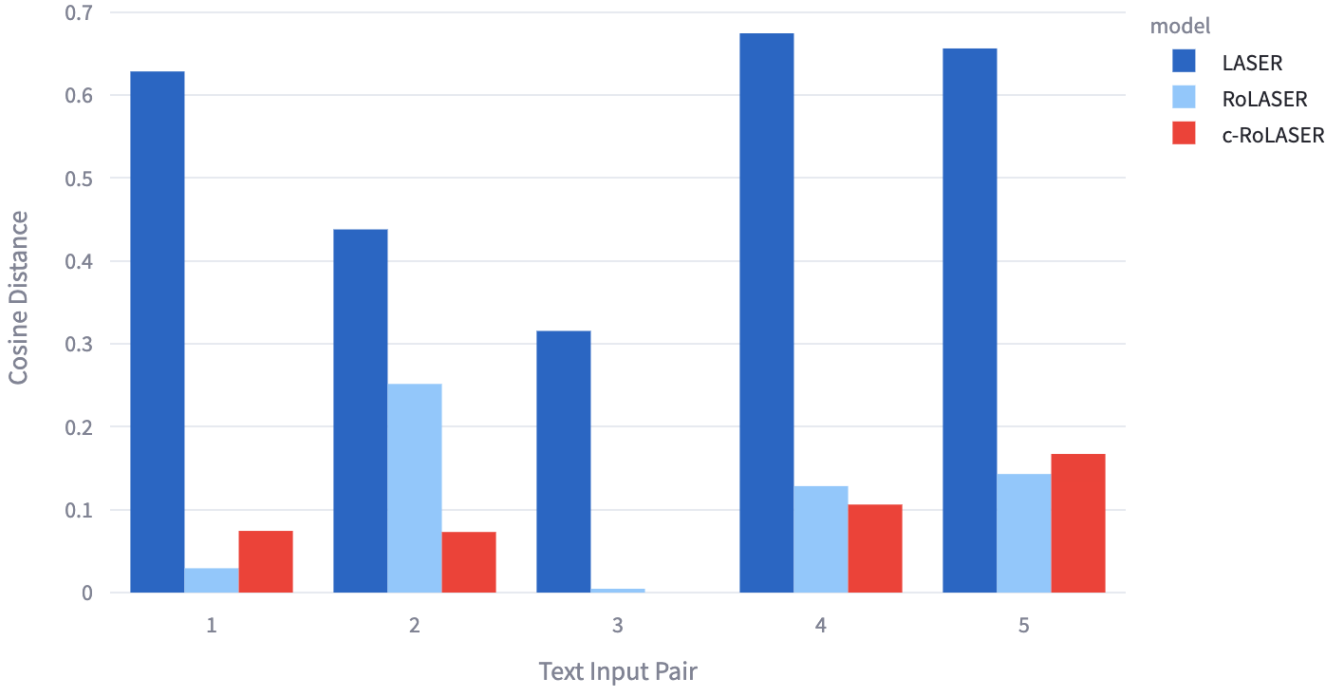
Evaluation on downstream tasks (2)



V. Conclusion

(c-)RoLASER's UGC embeddings

Standard text 1:	Non-standard text 1:
See you tomorrow.	See you t03orro3.
Standard text 2:	**Non-standard text 2:**
See you tomorrow.	C. U. tomorrow.
Standard text 3:	**Non-standard text 3:**
See you tomorrow.	sea you tomorrow.
Standard text 4:	**Non-standard text 4:**
See you tomorrow.	See yo utomorrow.
Standard text 5:	**Non-standard text 5:**
See you tomorrow.	Cu 2moro.



RoLASER Demo App

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

Thaaanx!!!

Do u hv any qweschuns?

Paper

https://aclanthology.org/2024.lrec-main.958/

RoLASER Demo App

https://huggingface.co/spaces/lydianish/rolaser-demo

Github

https://github.com/lydianish/RoLASER