## Motivation

**Machine Translation** for low resource languages has low performance largely due to **lack of training data.**

**Back-translation** relies on high quality monolingual data.

**Monolingual data** is “locked” in PDFs & images.

**OCR models** can “unlock” it but there is no comprehensive **evaluation of their performance and their impact on MT**.

## OCR Benchmark

- **Real life PDFs from UDHR**
  - Translated into over 500 languages
  - Most languages have both PDF and Text
  - Each document has 30 short articles, on average 3 sentences each
- **Artificially created PDFs: Flores 101**
  - 3,001 sentences, English Wikipedia
  - 101 languages, a wide variety of domains

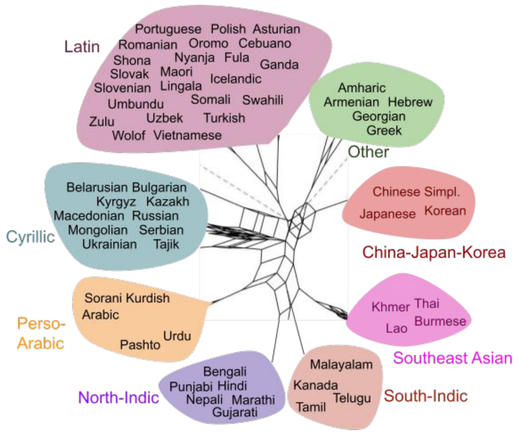


Fig. 1: We select **60 languages** that are in both datasets and prioritize **low resource languages** with low resource **scripts**

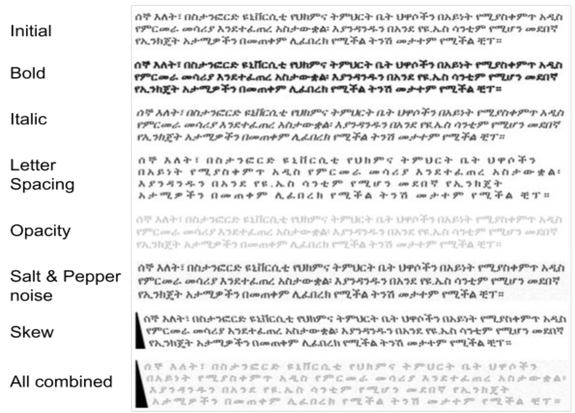


Fig. 2: **Data augmentation** sample on Amharic artificial PDF from Flores 101

# OCR Improves Machine Translation for Low-Resource Languages

*Oana Ignat, Jean Maillard, Vishrav Chaudhary, Paco Guzmán*

Contact: oignat@umich.edu

1. Novel benchmark of real and synthetic data, enriched with noise, 60 low-resource languages
2. Evaluate state-of-the-art OCR systems on our benchmark and analyse results based on script and location
3. Measure OCR impact on Machine Translation (MT):
   - OCR monolingual data can increase MT performance
   - Downstream impact of OCR errors in back translation

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

**Data and Code:**
https://github.com/facebookresearch/flores

## OCR Evaluation

- *Google Vision API* > *Tesseract*
- *Flores 101* easier than *UDHR*
- *UDHR synth* easier than *UDHR* (diff. in performance due to visual info. and not content)

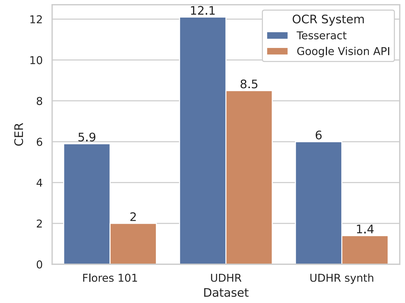


Fig. 2: Average CER (the lower, the better) of the **SOTA OCR systems, across datasets, over 60 languages.**

- Artificial data easier to recognize
- *Latin&Cyrillic* best performance
- *Perso-Arabic* performs badly
- Performance varies per language/ type of data: *North Indic, South Indic, SEA & Other* good on artificial but poor on real data

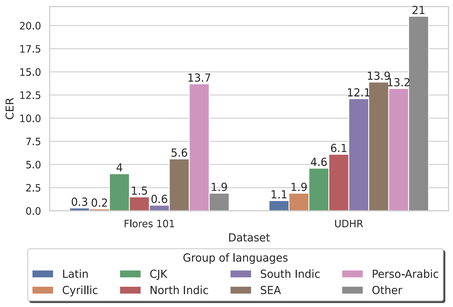


Fig. 3: Average CER (the lower, the better) of **best performing OCR model, across groups of languages**

## OCR Impact on MT

- **Performance** of the SOTA MT model *M2M-124* increased significantly when fine-tuned on OCR-ed data

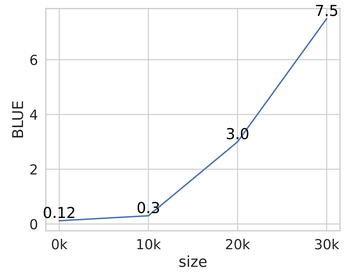


Fig. 4: **English to Nepali MT**, fine-tuning *M2M-124* on OCR-ed *Nepali books corpus.*

Back-translation on Khmer, Pashto & Tamil


- Translation quality is robust to small noise
- Replacements most damaging
- More data leads to higher/ more rapid performance decrease

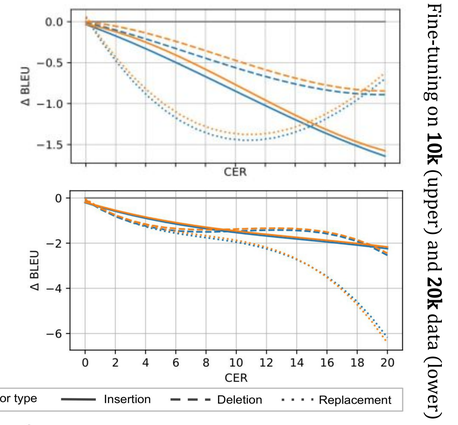


Fig. 5: **OCR errors impact on MT performance:**
ΔBLEU (*M2M-124* fine-tuned on OCR-ed data, *M2M-124* pretrained)
ΔBLEU (*M2M-124* fine-tuned on OCR-ed data, *M2M-124* fine-tuned on real data)

Fine-tuning on **10k** (upper) and **20k** data (lower)