EdinHelsOW WMT 2025 CreoleMT System Description: Improving Lusophone Creole Translation through Data Augmentation, Model Merging and LLM Post-editing

Paper

Jacqueline Rowe, Ona de Gibert, Mateusz Klimaszewski, Coleman Haley, Alexandra Birch, Yves Scherrer

We present our submissions to the WMT 2025 Creole Translation Shared Task. For seven Portuguese-based Creole languages, we augment the Task datasets with extra parallel and monolingual data in four languages. We leverage a range of low-resource MT techniques including back-translation, data distillation and model merging, as well as cross-lingual transfer from Portuguese data and embedding weights.

BACKGROUND

- Creole languages arise through contact between two or more languages from different families over a short space of time. They are often extremely **low-resource**, with few datasets, NLP tools or translation systems available.
- The 2025 Shared Task on Machine Translation for Creole Languages uses datasets and baseline models (**KreyòlMT** and **CreoleVal**) from two recent works covering over 50 Creole languages between them.
- For our submission, we focus on translation between **seven Portuguese-based Creoles** and English (eng):
  - Six Creoles are spoken in West and Central Africa (right) and one is spoken in the Dutch Caribbean (left).
  - pov, pre, aoa, cri and fab are extremely low-resource, with 100-300 parallel sentences provided by the organisers
  - kea has 1.5k organiser-provided sentences and pap has 67k. Both have featured in other MT and NLP modelling efforts.

SYSTEM DESIGN

Collect Data (Parallel, Monolingual and bilingual lexicons)

→ **Backtranslate** monolingual data ⇄ **Forward translate** some parallel data

→ **Combine, clean and re-split datasets**

→ **Finetune (FT) 3 pre-trained multilingual translation models**: KreyòlMT (KMT), mBART, NLLB

→ **Find best FT setup for each base model**:
- which size of base model?
- m2m, m2o, o2m — FT for specific translation directions?
- use distilled data instead of normal data?
- with high-quality Portuguese data in training?
- initialise Creole weights with Portuguese embeddings?

evaluate using best checkpoints per language, as per validation sets

→ **Merge** best checkpoints together

→ Prompt LLMs to **post-edit** system outputs, including bilingual lexicons in prompt context (OpenAI, Mistral AI, Gemini)

→ **Select best models for submissions:**
1. Best XX-eng/eng-XX models
2. Best models per language pair (LP)
3. Best post-edited systems

DATA

Numbers of parallel sents with eng in each Creole*

pap	110k
pov	46k
kea	12k
cri	2k
pre	0.2k
fab	0.1k
aoa	0.1k

*including all train/val/test splits and both KreyòlMT data and our data (including back-translations).

RESULTS

1. Different base models and FT settings benefit different LPs

Fig 1: Performance of different fine-tuned base models on pap (top) and aoa (bottom) KreyolMT test set. Compare results for basic setup, initialising Creoles with Portuguese embeddings, adding por data and using distilled data. KreyòlMT model baseline in red.

2. Model merging and LLM post-editing with lexicons boosts performance on some LPs

Fig 2: Results of our best finetuned model checkpoints (FT), our best merged models, and our best post-edited (PE) system outputs on high-resource (top) and low-resource (bottom) LPs on KreyòlMT test set.

3. Utilising alternative test sets

None of our models beat the KreyòlMT baseline model scores on the KreyòlMT test set for kea, pov or cri, and it became clear that that this was due to dataset contamination. We re-evaluated some of our submitted models on decontaminated test sets and beat the KreyòlMT baseline model on all LPs.

Fig. 3: Our models and baseline KreyòlMT model evaluated on a decontaminated test set. Submitted models are best finetuned or merged models for (1) eng-XX and XX-eng directions and (2) per LP (as evaluated on original KMT test set).

This research was funded in part by the UKRI AI Centre for Doctoral Training in Responsible and Trustworthy in-the-world Natural Language Processing (grant ref: EP/Y030656/1).