Leveraging Portuguese Latent Spaces for Kriolu to English Translation with Monolingual Language Model Priors

Kevin Santiago   Matthew Wilmot   Peter Lin

THE ISSUE OF CAPTURING LOW RESOURCE ORTHOGRAPHY

Cape Verdean Creole (Kriolu) is an oral, low-resource dialect of Portuguese with highly variable spelling and very little parallel data for accurate machine translation (MT)

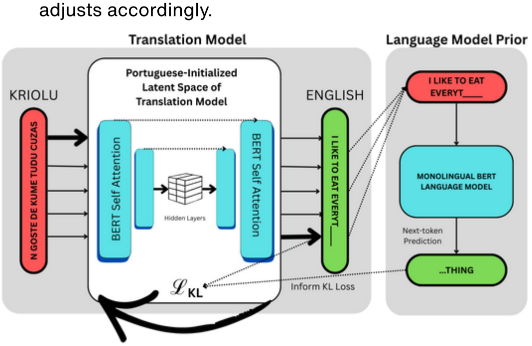
As a result, we noticed that most generalizable low-resource MT methods still treat all languages as independent, so they fail to fully capture the non-standard orthography of a language like Kriolu due to the lack of data for these models to fit correctly.

We address this by exploiting Kriolu's Portuguese roots, pre-embedding the MT encoder with different **Portuguese/Brazilian BERT** latent spaces (plus matching monolingual LMs) and comparing which initialization gives a better start for capturing Kriolu's structure.

PORTUGUESE-INITIALIZED BERT TM WITH MONOLINGUAL ENGLISH LM PRIOR

We build upon the work of Baziotis et. al. and use a Translation Model (TM) with a Language Model (LM) prior setup

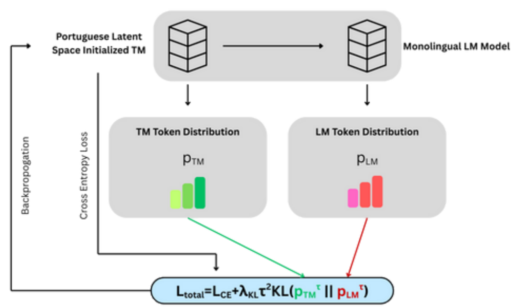
- **Portuguese Initialized TM:** Swappable TM block with Portuguese trained weights
- **English LM Prior during training:** Given the gold English sequence, this LM produces its own token distribution
- **Combined objective using Kullback–Leibler (KL) Loss**: KL loss between TM and LM logits accounts for disagreement between the TM and LM outputs and adjusts accordingly.



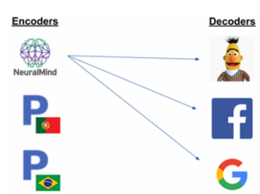
FINDING THE BEST LATENT SPACE AND LM COMBINATION

Goal: Force a TM to properly fit Kriolu to English translation using two priors upstream and downstream.

- **Upstream:** Capture fundamental orthographical nuances that Kriolu inherits from its root language (Portuguese).
- **Downstream:** Nudge TM output towards high-probability English using KL Loss.



To find the best combination of these upstream and downstream priors we tested all combinations of the following latent spaces and monolingual English LMs



Latent Spaces

- Bert-base-Portuguese
- Albertina-ptpt
- Albertina-ptbr

English LMs

- Bert-base
- RoBERTa-base
- Electra

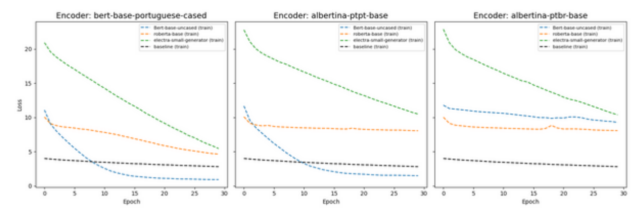
The **Kreyol-MT** Dataset was used for training and evaluation (filtered for Kabuverdianu)

Sentence	Cape Verdean Kriolu	English
S1	Dipôs di xinti konfortável ku formatason y edison di web, más tardi, bu pode kria bu própi website.	After you become comfortable with formatting and editing on the web, then later, you might create your own website.
S2	Es sabe ma na kel tenpu, es tinha ses kazinha.	They know that during that time, they had their house.
S3	Nos povu tenba ki vense, nos téra tenba ki liberta, nos ómi ku mudjer tenba ki vive na liberdádi, na pás y na pugrésu.	Our people had to win, our country had to liberate itself, our men and women had to live in liberty, peace and progress.

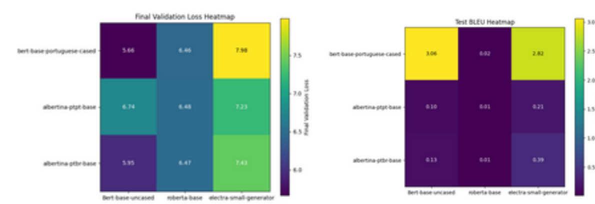
RESULTS

Training: Over the course of 30 epochs

- **bert-base decoder** models and latent spaces trained better (averaging final loss of 1.2).
- **bert-base-portuguese encoder/bert-base decoder** had the lowest validation loss
- **All models** outperform baseline validation loss of 8.69



Improved Bleu Scores: Our method saw an improvement on the baseline Bleu score (0.08) for 6 out of 9 encoder-decoder pairs.



Baseline: Randomly initialized seq2seq latent space with Bert-base English LM as prior

REFERENCES

[1] Christos Baziotis et al. 2020. *Language Model Prior for Low-Resource Neural Machine Translation* (EMNLP 2020)

[2] Nathaniel R. Robinson et al. *Kreyòl-mt: Building mt for latin american, caribbean and colonial african creole languages* (ACL 2024)

[3] Kishore Papineni et al. Bleu: a method for automatic evaluation of machine translation. (ACL 2002)

github.com/mcgwilmo