Model Training Documentation

Overview

This document outlines the training process of a neural machine translation model for Dyula to French. The model was trained using the JoeyNMT framework, with several versions (from version 1 to version 19) reflecting progressive adjustments to hyperparameters, architecture modifications, and data preprocessing techniques. The project's primary aim was to enhance translation accuracy, leveraging various approaches, including back-translation data.

Data

The dataset used for the initial model training included standard parallel corpora for Dyula to French. Back-translation data generated from Meta's NLLB (No Language Left Behind) model was incorporated. NLLB model generated synthetic Dyula data by translating French sentences back into Dyula, enriching the original training set.

Training Process

Versioning and Hyperparameter Tuning

1. **Version 1**: The first version of the model was trained using the `uvci/Koumankan_mt_dyu_fr` dataset. This version utilized default hyperparameters from JoeyNMT's transformer architecture, with slight adjustments to accommodate the dataset size.
2. **Versions 2 to 5**: Adjustments were made to the learning rate, optimizer type, and batch size. These versions aimed to strike a balance between overfitting and underfitting by tuning dropout rates and other regularization techniques.
3. **Version 6 to 10**: These versions experimented with different embedding sizes and model depth (number of layers).
4. **Version 11 to 15**: Aggressive data augmentation techniques were applied, including noise injection into the input sentences during training, and back-translation data generated using Meta’s NLLB 1.3B parameter model. The learning rate schedule was altered.
5. **Version 16 to 19**: In the final versions, the focus shifted to fine-tuning the model for higher accuracy on the test set. Techniques such as beam search decoding were applied to improve translation fluency and adequacy. These versions also utilized the final, optimized hyperparameters, as determined by grid search over the earlier versions.

Back-Translation Data from NLLB

In the mid-training phase (versions 11 to 15), data from Meta's NLLB model played a key role. This approach generated Dyula sentences by translating from French back into Dyula, which was then used to enhance the original training set.

The workflow includes:

1. Loading the NLLB pre-trained model
2. Translating the French source sentences back into Dyula
3. Post-processing the translated output to ensure quality before integrating it into the training pipeline.