Eleftheria Bargiota

Personal Project No.2

2025

1 Text Normalization

The main goal of this project is to propose a solution that normalize given text information, remove unecessary data and keep only important information in the output. The given dataset is a csv file consisting of writer's name and other information and the normalization version of it. For this task my suggestion is the utilization of a Seq2Seq (Sequence-to-Sequence) model is a type of neural network architecture designed to convert one sequence of data into another such the goal of this project. Seq2Seq has been widely used before for tasks like Language translation and Text summarization.

1.1 Code Implementation

The project was implemented in Python using Google Colab Pro+, utilizing 50GB of RAM. The

1. **spacy library**: In order to be able to expand text normalization in non-latin characters we used *spacy* which is an open-source library for Natural Language Processing. The languages I included are English, Japanese, Chinese, Arabic and Russian. The github link of this project is below:

   GithubLink

2. **Tokenize** : For tokenization we used the function word-tokenize from NLTK package, to break down the text in smaller units.

3. **Vocabulary** : **SRC**(source) defines how to preprocess the input data (e.g., tokenization, adding special tokens like ¡sos¿ and ¡eos¿, etc.) for the source text.and **TRG**(target) defines how to preprocess the output data for the target text.

4. **Dataset**: The dataset was split in 3 subsets. Training set consists of 5000 examples, and validation and test set consists of 2500 examples each.

5. **Architecture**: The Seq2Seq architecture consists of an Encoder Class that processes the input sequence and pass it to the decoder, and a Decoder class that generates the output sequence based on the context vector passed from the encoder (generates one token at a time).

6. **Normalization**: For the normalization we implemented th *normalize-sentence* function. The purpose of the method is to take a raw sentence, tokenize it, and then generate a normalized version of the sentence using the Seq2Seq model.

7. **Training**: We trained the Seq2Seq model for 15 epochs.The training loss is decreasing, which means the model is learning to map the source sequence (raw writer's text) to the target sequence (normalized text). The validation loss also decreases, which is a good sign. It indicates that the model is not just overfitting to the training data but is also generalizing well to unseen data. We stopped the training in the 15 epochs since after 18 epochs the validation loss started increasing and the model would propably overfit to the data.

8. **Evaluation** BLEU score