I. Project Proposal

Problem Statement

There are plenty of machine translators, some with high accuracy, and others with much needed improvement. I want to build a machine translator with high accuracy, starting with english-french, to eventually build models to translate less popular languages such as Amharic.

Client(s)

An english speaker interested in learning french or a french speaker interested in learning english. The end goal is to develop an app, so users can translate easily on the go.

The Data

The data was acquired from manythings.org which has a list of bilingual sentence pairs, sentence pairs between english and many other languages. The dataset contains 179,009 sentence pairs between english and french.

Outline of problem solving method

- Download the zip file from manythings.org and create a dataframe with 2 columns, english and french, containing the english and french sentence pairs.
- Explore the data to see if there are any missing values or outliers
- Apply data preprocessing and tokenization
- Data visualization to gain some insight by plotting distributions of word/sentence behaviors for example.
- Apply some statistical analysis
- GRU &/or LSTM modeling using PyTorch

Deliverable(s)

- Milestone Report
- Final Report & Slide Deck
- Code in Jupyter notebooks