

# Seq2Seq Translation

Goal: Constructing a sequence-to-sequence (encoder-decoder) model for machine translation

Write a Python program in the form of ipynb, following the details below.

Using PyTorch is recommended.

- Due: 2024-05-13 (Mon) 23:59
- Submit on eTL

## [1] Dataset

- A bilingual dataset from <http://www.manythings.org/anki/>
- You can choose any dataset of the given list.
- Don't use small datasets (less than 5,000 examples)

## [2] Task

- A machine translation task
- Input (source): English sentence
- Output (target): The language that you selected from the list

## [3] Elements that you should include in the program

- Construction of data iterators (and data preprocessing if needed)
  - **Please include the codes/cells that download and uncompress the data file in the program.** (so that we can simply run the program)
  - You'll have to split the dataset into training and test sets. Set the ratio of training:test = 9:1.
- The Python Classes for Encoder, Decoder, and the Seq2Seq model
  - **No Transformers** for the encoder or decoder
- The training, evaluation processes of the constructed model
- Print the (average) BLEU score(s) as well as the loss values during model evaluation (on the test set).
  - You can use existing Python libraries for the BLEU score.
- A code snippet or a function for inference
  - A test sentence for inference: "He is a baseball player ."
  - Use the trained Seq2Seq model and print the translated sentence.
- Set and use appropriate values or methods for other elements that are not instructed.