Datasets

An overview of some useful datasets to train a seq2seq model with.

A. Seq2seq datasets

With the encoder-decoder model architecture, we have a model template for any sequence to sequence NLP task. Of course, each seq2seq task will require some tweaks to the model, but overall the architecture largely stays the same across all domains.

The most important part to creating a good seq2seq model is finding a lot of useful training data. For machine translation, there are tons of training datasets online. You can find many of the largest machine translation datasets here.

For training a dialog system such as a chatbot, a couple good datasets to use are the Ubuntu Dialogue Corpus and the Cornell Movie Corpus.

For text summarization, a good training set to use is the CNN/Daily Mail summarization dataset. Instructions for how to obtain the data can be found here.