Introduction

Our motive for this Wikipedia simplifier is converting Wikipedia articles into simple English so that they can be accessible to more and more people. And nonnative English speakers and also a person with some learning disability can also understand those articles by means of using limited vocabulary.

Model

So the basic idea behind this proposed model is based on repeated sampling by a pre-trained transformer based encoder that can generate meaningful sentences in the context of the text. To provide the context of previous sentences we will use the seed method (which means we will pass the previous sentences as a seed to pass on the context)
And then we will pass these sentences through a rejections pass filter(RPF) that will reject the sentences which fail to match the embedding score of the source; this will effectively control the output content. We can use a threshold as a minimum similarity parameter to configure RPF that way we can filter out between the repeated samples that we will generate for each input sentence.
This process for filtering will be adjusted by itself by changing t (hyperparameter for minimum threshold). And if no sample passes that condition, we will send the original sentence to the output unchanged and pass on two next sentences.

As an evaluation criterion for text simplifying we will use Flesch Kinkaid's readability score that is based on sentence length and word length. And for our dataset, we will use a simple Wikipedia text corpus by doing some preprocessing like converting it into sentences and then each sentence into some lexemes (sentence parts which can be separate). We will use a GPT-2 based transformer and google sentence encoder and other pipeline components (lexeme processor, pass filter, Kinkaid function for evaluation, etc) we will build.

We have many hyperparameters that will be changed accordingly to best optimize the model. T is the minimum threshold to pass through the rejection filter. {MIN, MINSOFT, MAX} are the unconditional minimum, punctuation marked minimum and maximum lexeme size in words.

This model has an average of 4.712% increase in the Flesch score which is very good for a completely independent machine learning model.