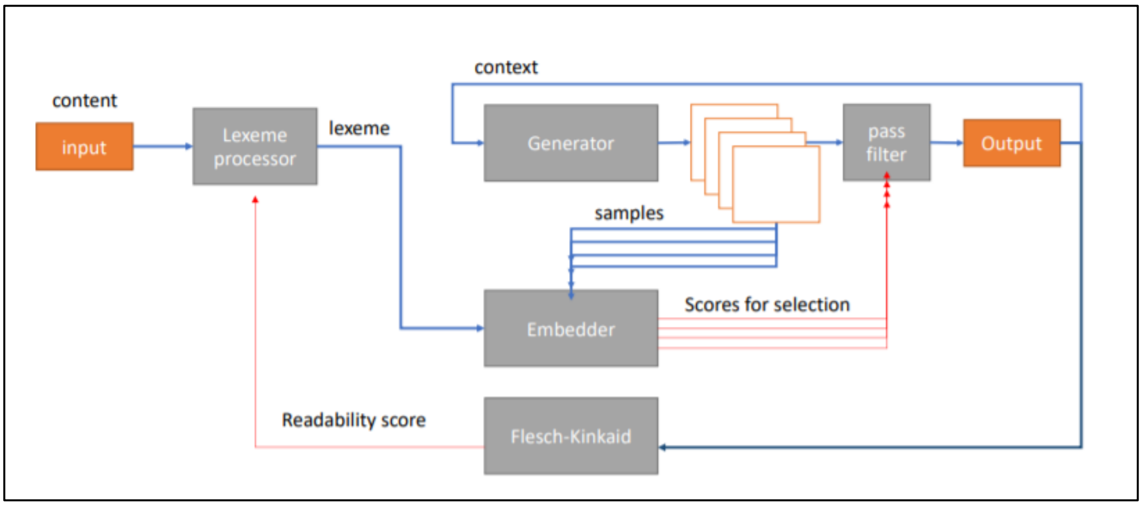
Wikipedia Simplifier: Summary

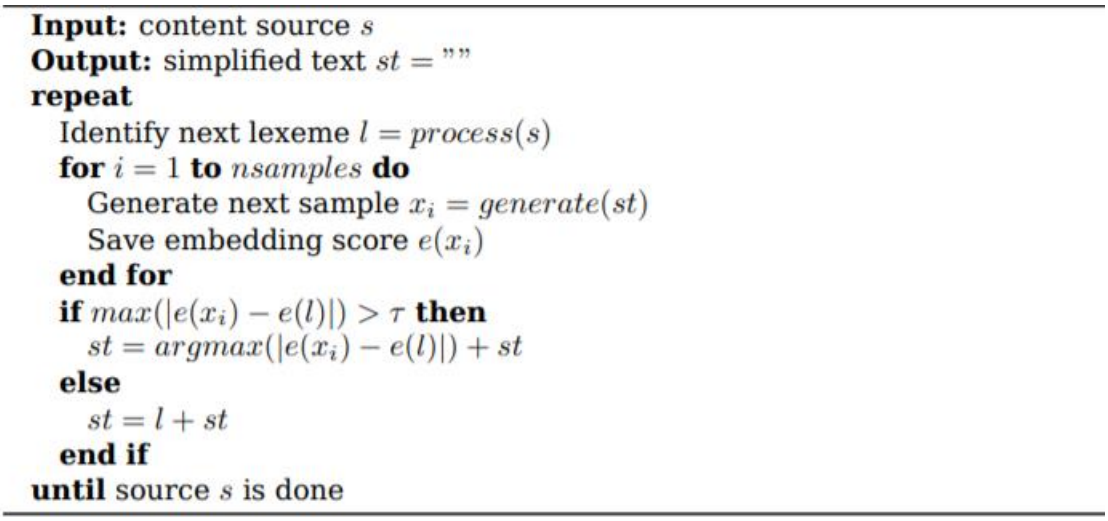
Aim – to simplify the Wikipedia articles with simplified vocabulary and sentence structures to make them more readable.

Approach – *GPT-2 transformer* was used for text generation. This pre-trained generator was fine-tuned on the Simple English Wikipedia articles. The model works as follows :

It takes in the source content. The source content is passed to the *Lexeme processor*, which splits the sentences into lexemes. Taking one lexeme at a time, this is given as a seed to the generator, which produces a set of possible samples in simple English that can replace that lexeme. This lexeme and the generated samples are then passed into the *Embedder*, which converts these to embeddings that can be compared directly for measuring similarity. *Google Sentence Encoder* was used for this purpose. These embeddings are then passed to the *Pass-Filter*, which compares the original lexeme and the embeddings. It then chooses the generated sample that is closest to the lexeme. If its similarity is above the threshold value, then this sample is accepted for the simplified text; otherwise, the lexeme is accepted unchanged. This simplified text is passed as a seed to the generator for producing samples for the next lexeme in the source content. These steps are then repeated until the entire source content is simplified.

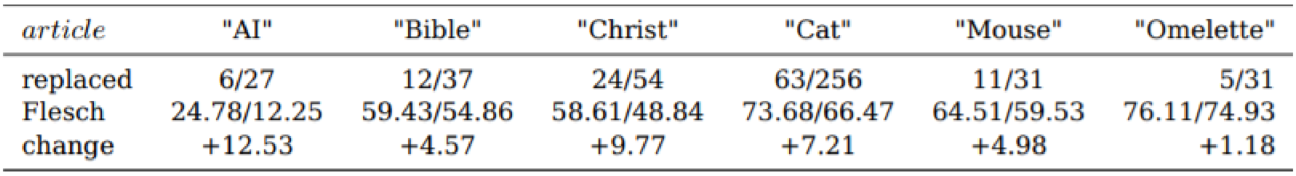


Algorithm –



```
Input: content source $s$
Output: simplified text $st = ""$
repeat
  Identify next lexeme $l = process(s)$
  for $i = 1$ to $nsamples$ do
    Generate next sample $x_i = generate(st)$
    Save embedding score $e(x_i)$
  end for
  if $max(|e(x_i) - e(l)|) > \tau$ then
    $st = argmax(|e(x_i) - e(l)|) + st$
  else
    $st = l + st$
  end if
until source $s$ is done
```

Result – This RPF framework was applied to twenty Wikipedia articles with their Simple English counterparts unseen in training for testing. In total, 556 lexemes were processed and 210 replacements were found (37.7% transformation rate). The average improvement in readability score was +4.712, with standard deviation 3.32.



article	"AI"	"Bible"	"Christ"	"Cat"	"Mouse"	"Omelette"
replaced	6/27	12/37	24/54	63/256	11/31	5/31
Flesch	24.78/12.25	59.43/54.86	58.61/48.84	73.68/66.47	64.51/59.53	76.11/74.93
change	+12.53	+4.57	+9.77	+7.21	+4.98	+1.18

Error Analysis –

1. Google Sentence Encoder was not always up to the mark. It sometimes failed to indicate a mismatch between phrases with different contexts.
2. For words that were not in the training set, the model produces the replacements, which looked similar but were not always factually consistent.
3. The framework sometimes oversimplified the sentences. It also picked unwanted language biases from the training set.
4. The model did not process the quotes differently, thereby giving replacement for them too. But this altered the quote, which was not intended.