INTRO. TO NLP - PROJECT - QUESTION ANSWERING TASK ON SQUAD DATASET

Arnau Ruiz (192961), Simone Gigante (205702), Enrique Torres (205203)

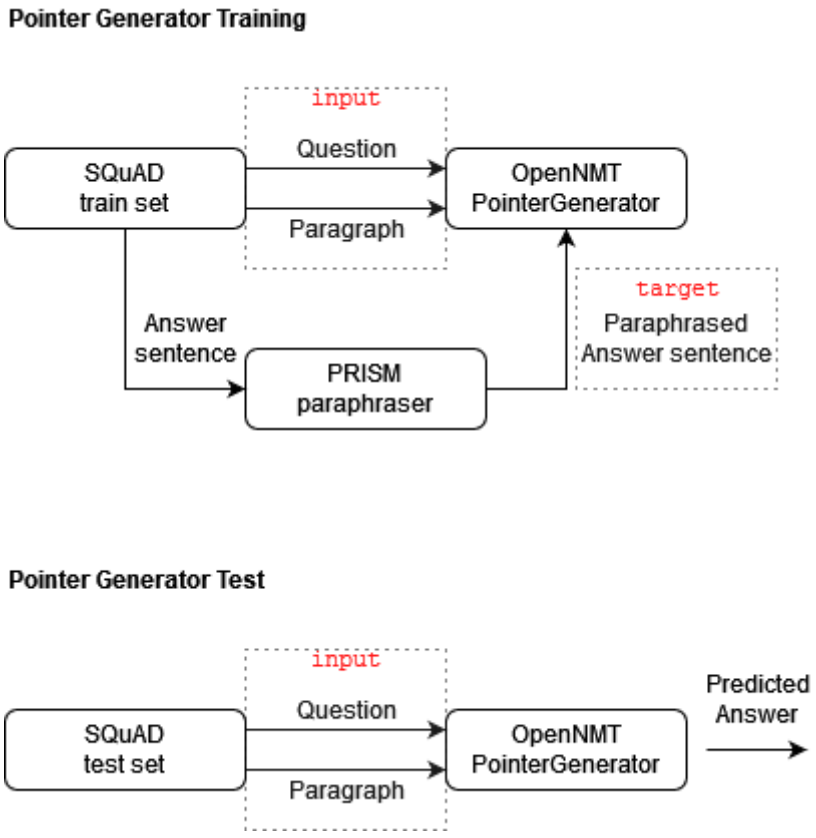
DESCRIPTION

Originally, this project consisted of performing a question-answering approach on SQuAD dataset by combining a paraphraser with a pointer generator model.

In a first attempt, we have used the prism model (https://github.com/thompsonb/prism) in order to paraphrase the sentences from which our target answers belong. We wanted our model to generate an answer that was not a copy from the paragraph.

We have provided sequences of questions and paragraphs of text as training data to a pointer generator model (https://github.com/OpenNMT/OpenNMT-py). By using paraphrased answer sentences as target data, we aimed to generate coherent answers that are like a summary or an abstract of the sentence where the answer is.

The following diagram summarizes our original idea:



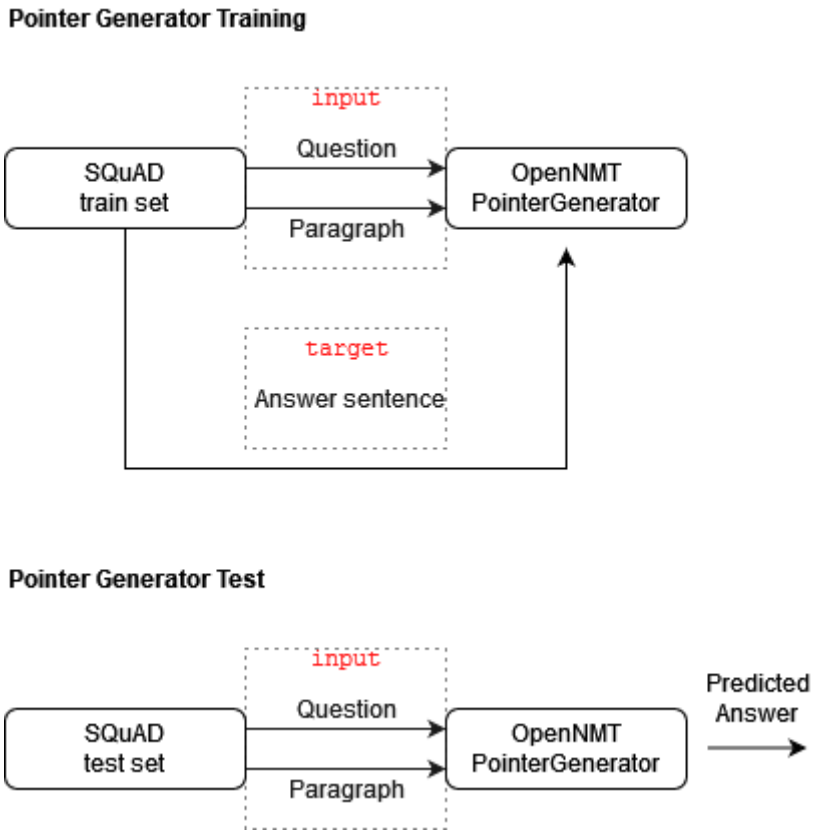
Unfortunately, this led to an unfeasible computational cost, given that Google Colab only provides us a limited use of the GPU accelerator.

For that reason we decided to not paraphrase target answer sentences and simply use current answer sentences as target examples.

INTRO. TO NLP - PROJECT - QUESTION ANSWERING TASK ON SQUAD DATASET

Arnau Ruiz (192961), Simone Gigante (205702), Enrique Torres (205203)

The following diagram summarizes the implemented idea:



Additionally, in order to check variations in the coherence of predicted answers, we have respectively trained and tested the pointer generator by using raw and preprocessed (alphanumeric characters only) text data. We have implemented this by providing a boolean parameter, **text_preprocess,** to *get_pg_data_samples(),* the function that creates the input text files for the model.

RESULTS

Predicted answers become much more coherent when using preprocessed data, in a way that using raw data makes these answers meaningless.

Using raw data:

```
[2021-12-08 17:40:48,706 INFO]
SENT 5: ['Who', 'upon', 'arriving', 'gave', 'the', 'original', 'viking', 'settlers', 'a', 'common', 'identity', '?',
PRED 5: Before ? the Frankish Frankish Frankish Frankish Frankish almost almost almost almost almost almost almost s
PRED SCORE: -194.3082
```

Arnau Ruiz (192961), Simone Gigante (205702), Enrique Torres (205203)

Using preprocessed data:

```
[2021-12-08 18:09:46,979 INFO]
SENT 5: ['Who', 'upon', 'arriving', 'gave', 'the', 'original', 'viking', 'settlers', 'a', 'common', 'identity', '?',
PRED 5: Before Rollo 's arrival its populations did not differ from Picardy
PRED SCORE: -1.5713
```

However, by looking at preprocessed data predictions, most of the predicted answers reference some piece of the text of the paragraph. So for most cases, we have seen that the model is not able to clearly refer to the question.

```
[2021-12-08 18:09:47,164 INFO]
SENT 27: ['What', 'ruined', 'Richard', "'s", 'plans', 'to', 'reach', 'Acre', '?', 'In', 'April', '1191', 'Richard', 'the', 'Lionhearted', 'left', 'Messina',
PRED 27: In April 1191 Richard the Lionhearted
PRED SCORE: -1.9611
```

A possible explanation for this is that we are not correctly approaching the question answering task with the OpenNMT model, which is currently set to provide titles (summary sentences) to input sequences of text.

In other words, we have not done anything special to make the model “realize” that it has to find information in the paragraph part related with the question part in the sequence. It works as usual, taking the entire sequence and giving a title as output.

Given that the question occupies a small part of the entire sequence, and probably does not contain highly representative syntactic or semantic content, the model tends to verbatim summarize portions of the paragraph text in the sequence.

However, there are also some examples of good (coherent) question answering predictions:

```
[2021-12-08 18:09:49,177 INFO]
SENT 46: ['What', 'kind', 'of', 'problems', 'are', 'one', 'of', 'the', 'main', 'topics', 'studied', 'in', 'computational', 'complexity', 'theory', '?', 'Decision',
PRED 46: Decision problems are one of the central objects of study in computational
PRED SCORE: -2.3023
```

To be honest, we must say that these good examples tend to appear when the answer is in the first sentence of the paragraph, also the same first sentence in the input sequence after the question mark.

CONCLUSION

We have noticed that the performance of question answering task with this model is limited because, even though we have introduced questions in the input sequence, we are probably still using it as a summarizer.

However, we have found good examples in terms of coherence when the question answer is closely related to the summary meaning of the entire paragraph.