Abstractive Synopsis Generation

Github Link:
https://github.com/ReuelSam/DataMiningProject

NLP Project
➢   Reuel Samuel Sam (2018103053)
➢   Sayf Zakir Hussain (2018103059)

Introduction

This project focuses on the generation of a synopsis for a given passage. To do so, this project attempts to adopt the important Natural Language Processing task of Text Summarization. Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning.

Text Summarization follows 2 methodologies:

- Extractive Summarization
    - The important phrases and sentences are extracted as is from the text
- Abstractive Summarization
    - New sentences are generated as the summary for this given text

This project attempts to provide text summarization for news headlines taken from the InShort dataset
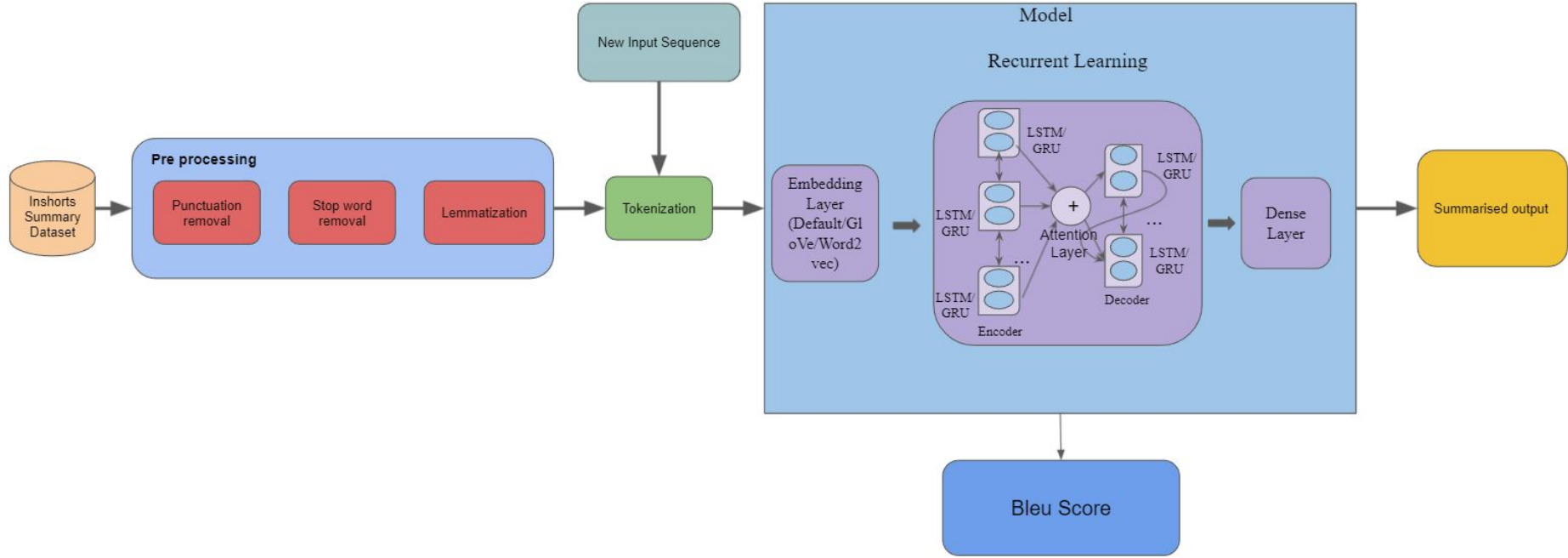
Objective and Dataset

The main aim of this project is to provide Abstractive Text Summarization for news articles through the use of LSTM based Encoder and Decoder structure.

For training purposes, the InShorts News Article Dataset has been identified. Inshorts is a news service that provides short summaries of news from around the web. This dataset contains headlines and summary of news items along with its source.

Dataset Link: https://www.kaggle.com/shashichander009/inshorts-news-data

Methodology Diagram

Performance

- Bleu Score:
    - BLEU, or the Bilingual Evaluation Understudy, is a score for comparing a candidate translation of text to one or more reference translations.
    - Although developed for translation, it can be used to evaluate text generated for a suite of natural language processing tasks.
    - Perfect BLEU Score = 1

$$Bleu \ = \ \rho.\, e^{\sum_{n=1}^{N} (\frac{1}{N}*\, log\, \rho n)}$$

where,

ρ = Brevity Penalty given by:

$$\rho \ = \ 1\, if\, c \ > \ r\, else\, e^{(1-\frac{r}{c})}$$

Performance

Model	Bleu Score
Model 1: BiGRU Vanilla (Paper Implementation)	0.5748
Model 2: LSTM Vanilla	0.5504
Model 3: LSTM with GloVe	0.6016
Model 4: LSTM with GloVe and Attention	0.7254
Model 5: LSTM with Word2Vec and Attention	0.7481
Model 6: GRU Vanilla	0.5320
Model 7: GRU with GloVe	0.6259
Model 8: GRU with Word2Vec	0.6208

Result: Best Model

```
In [48]:   print("Review:",seq2text(padded_xtest[13]))
           print("Original summary:",seq2summary(padded_ytest[13]))
           print("Predicted summary:",decode_sequence(padded_xtest[13]))
```

Review: rafael benitez appoint new manager relegation threaten premier league outfit newcastle united magpies sack previous manager steve mclaren early today benitez sack real madrid january seven month charge sign three year deal newcastle previously manage liverpool inter milan napoli valencia
Original summary: rafael benitez hired as newcastle united manager
Predicted summary: rafael benitez hired as newcastle united manager

```
In [49]:   print("Review:",seq2text(padded_xtest[25]))
           print("Original summary:",seq2summary(padded_ytest[25]))
           print("Predicted summary:",decode_sequence(padded_xtest[25]))
```

Review: mall china 39 s shanxi create rooster statue resemble us president elect donald trump small replicas statue available sale price start 3 800 11 lakh 32 foot tall statue notably new year china begin january 28 year rooster accord chinese zodiac
Original summary: china mall installs 39 donald trump like 39 rooster statue
Predicted summary: china mall installs 39 donald trump statue 39 china

```
In [50]:   print("Review:",seq2text(padded_xtest[48]))
           print("Original summary:",seq2summary(padded_ytest[48]))
           print("Predicted summary:",decode_sequence(padded_xtest[48]))
```

Review: usa become first nation register gold medal rio olympics 2016 19 year old virginia thrasher gold 10 metre air rifle day one event hold saturday thrasher finish ahead china 39 s du li yi sile silver bronze medal open china 39 s medal tally event
Original summary: usa wins first gold medal of rio olympics 2016
Predicted summary: usa wins first gold medal at rio olympics