Text Based News Summarization

Presenters: (Group 15)
Himani Shah (shahhi@iu.edu)
Samardeep Gurudatta (samgurud@iu.edu)
Mahadevan Iyer (mahiyer@iu.edu)

Text Based News Summarization

- Text summarization using Machine Learning is a classic problem with numerous applications in our daily lives.

- Multiple versions of the same news article are present and most of those articles are longer than 1000 words

- A solution that can condense news stories in a minimum number of words is urgently needed, making it simpler for individuals to ingest the massive amount of information published daily.

CNN-DAILY MAIL DATA

- id: a string containing the hexadecimal formatted SHA1 hash of the URL where the story was retrieved.
- article: a string containing the body of the news article
- highlights: a string containing the highlight of the article as written by the article author

(Sources: https://www.kaggle.com/gowrishankarp/newspaper-text-summarization-cnn-dailymail)

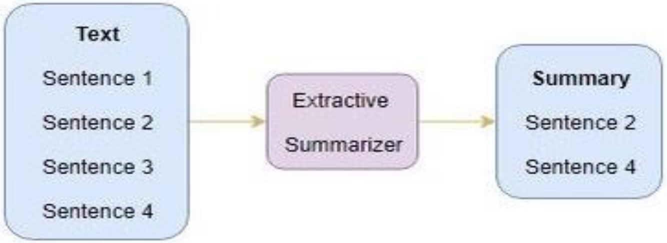
Related Work

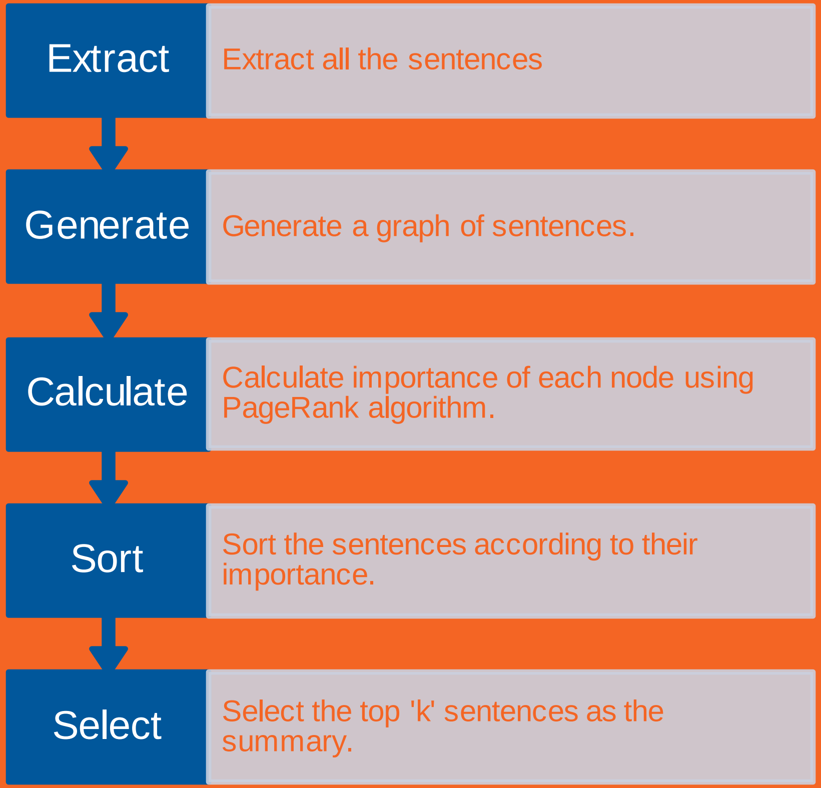
[1] S. R. K. Harinatha, B. T. Tasara and N. N. Qomariyah, "Evaluating Extractive Summarization Techniques on News Articles," 2021 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2021, pp. 88-94, doi: 10.1109/ISITIA52817.2021.9502230.

[2] Chen, V. (2017). An Examination of the CNN / DailyMail Neural Summarization Task.

[3] Nallapati, R., Zhou, B., Gulcehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Extractive Summarization using TextRank





Extract: Extract all the sentences

Generate: Generate a graph of sentences.

Calculate: Calculate importance of each node using PageRank algorithm.

Sort: Sort the sentences according to their importance.

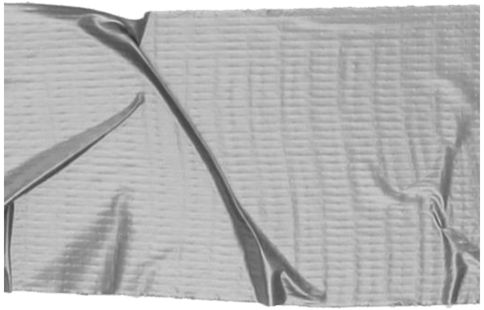
Select: Select the top 'k' sentences as the summary.

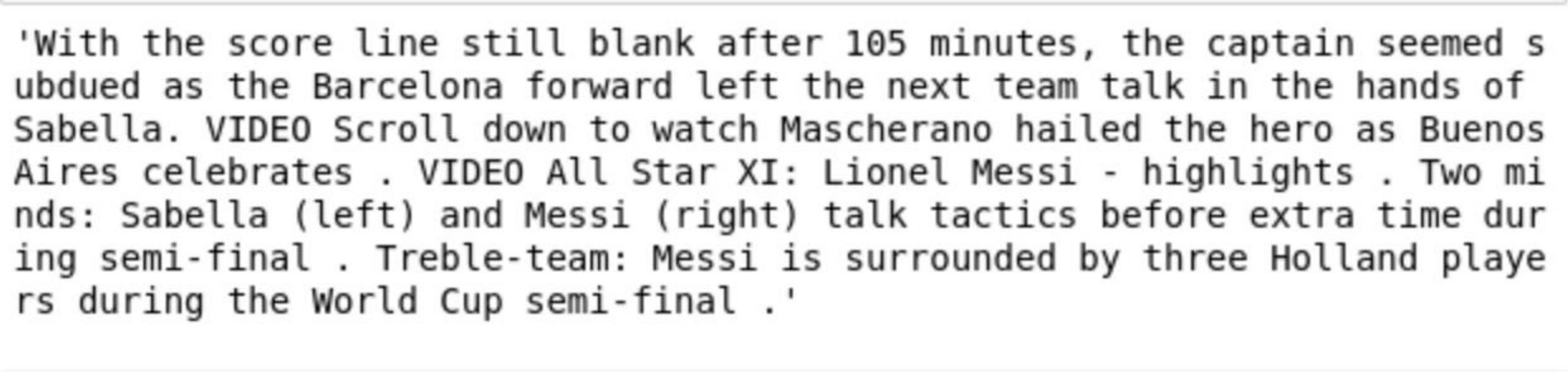
FOLLOWING ARE THE RESULTS OF EXTRACTIVE SUMMARIZATION USING TEXTRANK:

Precision	13%
Recall	26.3%
F1	16%

Evaluation Matrix:

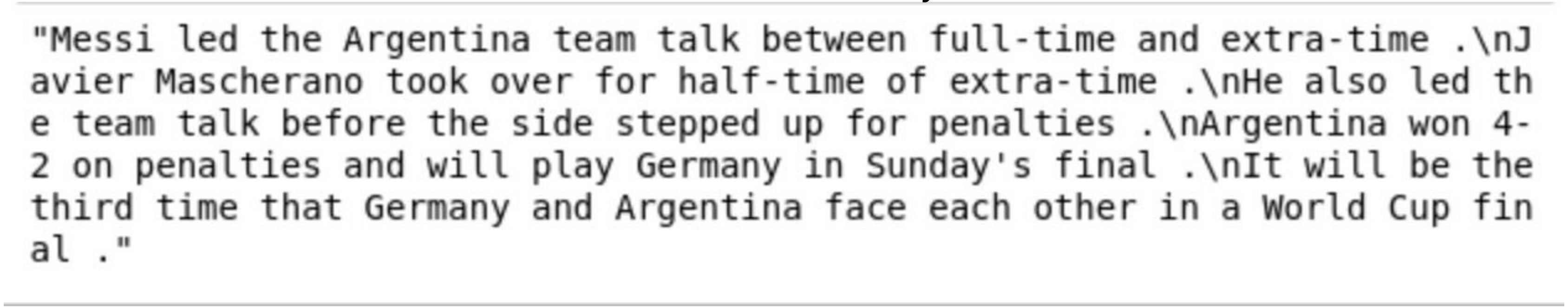
RougeL





```
'With the score line still blank after 105 minutes, the captain seemed s
ubdued as the Barcelona forward left the next team talk in the hands of
Sabella. VIDEO Scroll down to watch Mascherano hailed the hero as Buenos
Aires celebrates . VIDEO All Star XI: Lionel Messi - highlights . Two mi
nds: Sabella (left) and Messi (right) talk tactics before extra time dur
ing semi-final . Treble-team: Messi is surrounded by three Holland playe
rs during the World Cup semi-final .'
```

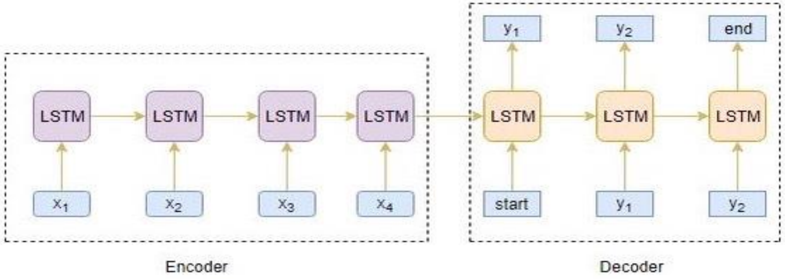
Predicted Summary



```
"Messi led the Argentina team talk between full-time and extra-time .\nJ
avier Mascherano took over for half-time of extra-time .\nHe also led th
e team talk before the side stepped up for penalties .\nArgentina won 4-
2 on penalties and will play Germany in Sunday's final .\nIt will be the
third time that Germany and Argentina face each other in a World Cup fin
al ."
```

Human Generated Summary

Abstractive Summarization using Encoder Decoder Architecture



Training

Inference

Bi-Directional LSTM

Model Parameters

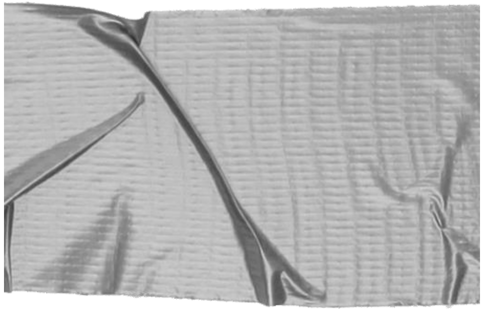
Word Embedding Dimension	300
LSTM Hidden Units	400
Max length of article	1250 words
Max length of summary	100 words
Optimizer	RMSprop (clipnorm=2.0)
Loss function	Sparse Categorical cross entropy
Batch Size	32/64

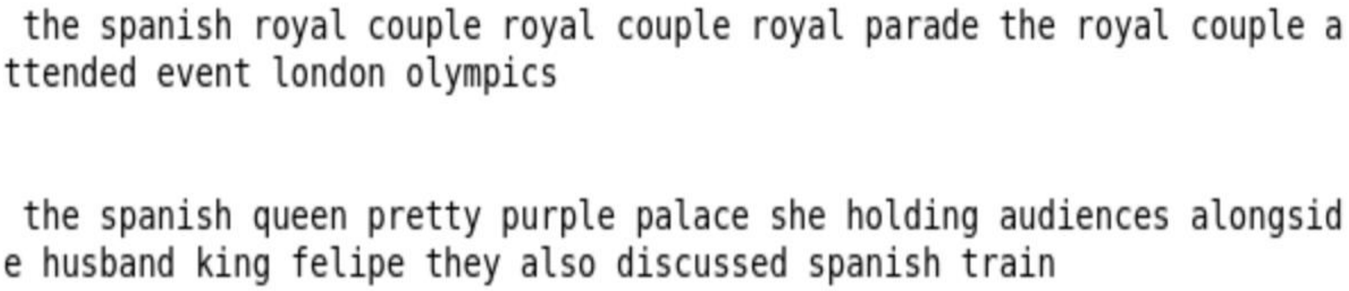
FOLLOWING ARE THE RESULTS:

Batch Size	32		64	
Learning Rate	0.001	0.005	0.001	0.005
Precision	13.9%	9.3%	13.6%	9.3%
Recall	10.4%	7.6%	10.9%	7.8%
F1	11.4%	7.2%	11.6%	8.1%
Val Loss	2.2274	2.2435	2.1879	2.2435
Epochs	10	8	12	8

Evaluation Matrix:

RougeL



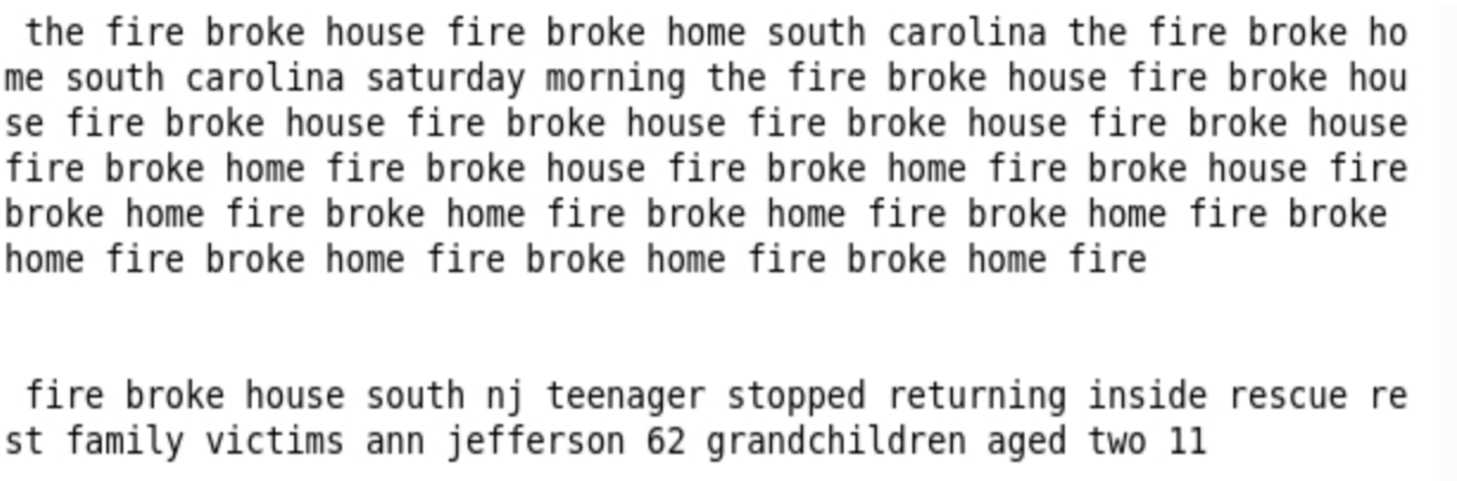


```
 the spanish royal couple royal couple royal parade the royal couple a
ttended event london olympics
```

Predicted Summary

```
 the spanish queen pretty purple palace she holding audiences alongsid
e husband king felipe they also discussed spanish train
```

Human Generated Summary



```
 the fire broke house fire broke home south carolina the fire broke ho
me south carolina saturday morning the fire broke house fire broke hou
se fire broke house fire broke house fire broke house fire broke house
fire broke home fire broke house fire broke home fire broke house fire
broke home fire broke home fire broke home fire broke home fire broke
home fire broke home fire broke home fire broke home fire
```

Predicted Summary

```
 fire broke house south nj teenager stopped returning inside rescue re
st family victims ann jefferson 62 grandchildren aged two 11
```

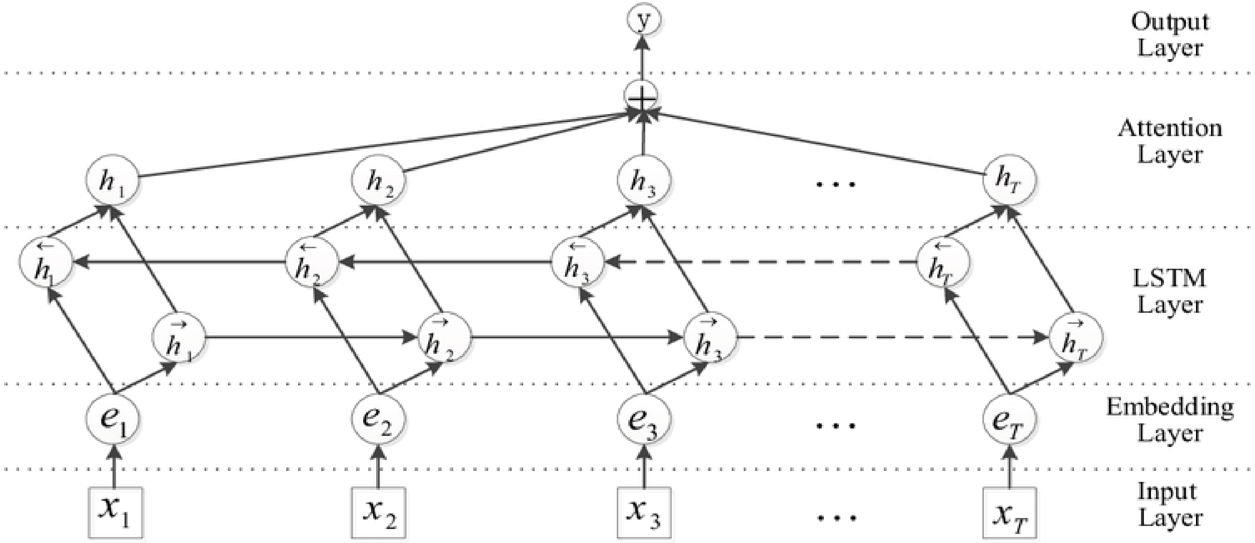
Human Generated Summary

Bi-Directional LSTM with Attention

Model Parameters

Word Embedding Dimension	500
LSTM Hidden Units	150
Max length of article	900 words
Max length of summary	70 words
Optimizer	RMSprop (clipnorm=2.0)
Loss function	Sparse Categorical cross entropy
Batch Size	64

Attention Mechanism



Bi-Directional LSTM with Attention

Model Parameters

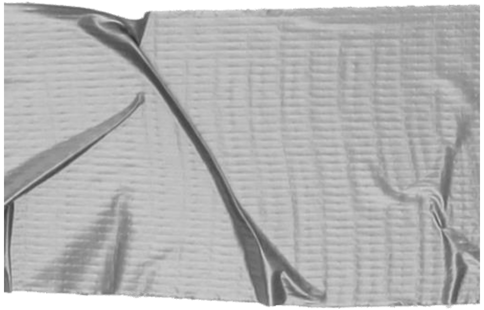
Word Embedding Dimension	500
LSTM Hidden Units	150
Max length of article	900 words
Max length of summary	70 words
Optimizer	RMSprop (clipvalue=1.0)
Loss function	Sparse Categorical cross entropy
Batch Size	128

FOLLOWING ARE THE RESULTS:

Batch Size	64		128	
Learning Rate	0.001	0.005	0.001	0.005
Precision	13.6%	11%	13.3%	14.1%
Recall	11.1%	9.3%	11.3%	11.4%
F1	11.7%	9.6%	11.8%	12%
Val Loss	3.0655	3.0440	2.9759	2.9170
Epochs	10	10	10	10

Evaluation Matrix:

RougeL



Key Takeaways

Answering: Research Questions 1 and 2

- The generated summaries are readable and make sense, however they contain repetitions and sometimes skip over important facts or get the plot wrong altogether.
- Seq2seq model without Attention Mechanism performs like Extractive Text Summarization when their RougeL scores are compared. Extractive summarizer had a higher recall than Seq2Seq model.
- Seq2seq model with Attention Mechanism performs slightly better than Seq2seq model without Attention Mechanism.

CONCLUSION

What have we Learned?

- How to use bi-directional LSTM for seq2seq modelling.

Challenges:

- We could only use 150k datapoints out of 280k due to high training time.

Difference between project proposed and final project:

- We dropped the idea unsupervised topic clustering as our model was able to capture context.
- The next step is to implement pointer generator model to improve our results.

THANK YOU