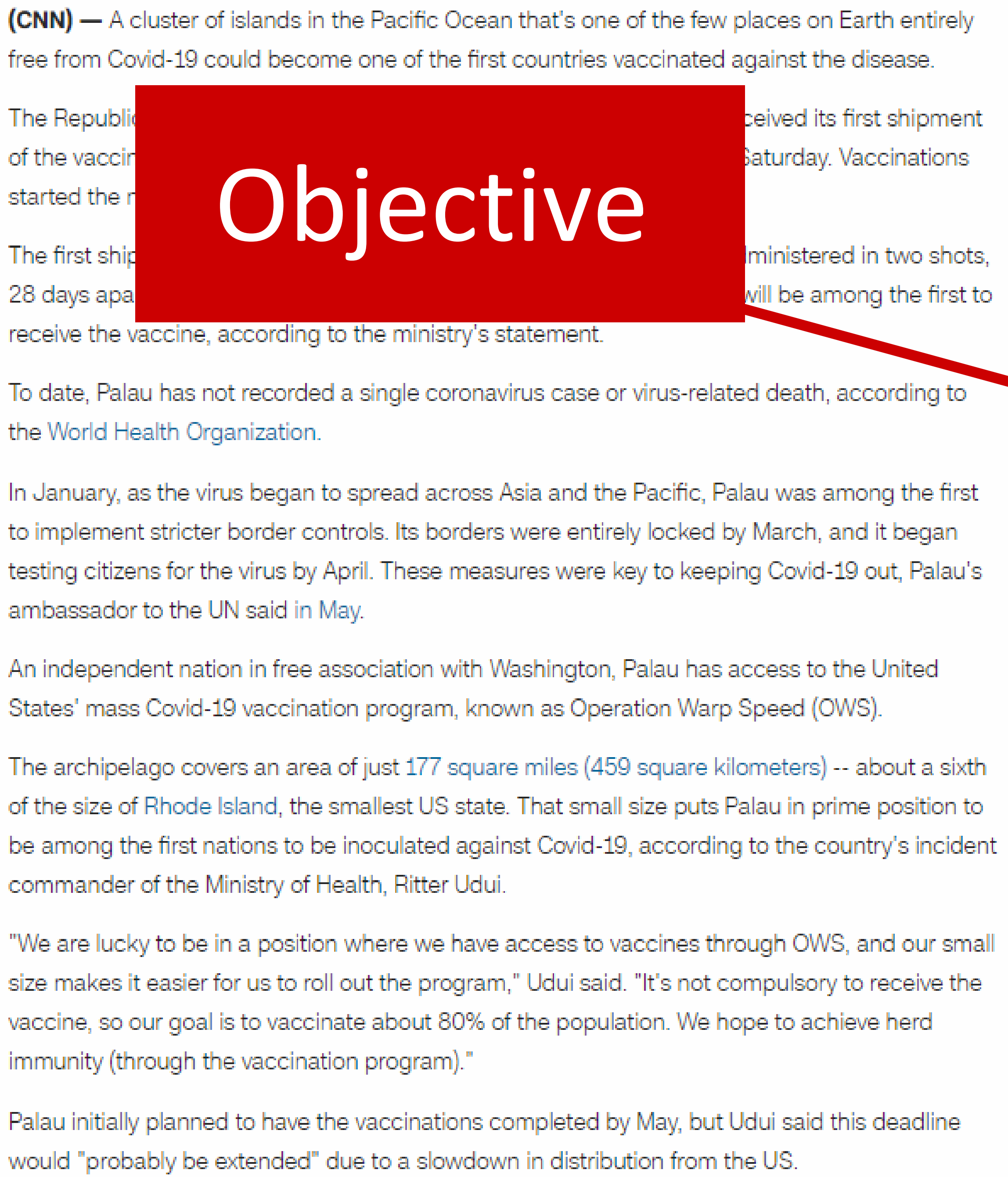
CNN STORY SUMMARIZATION

Christian Uccheddu - 800428
Federico De Servi - 812166



(CNN) — A cluster of islands in the Pacific Ocean that's one of the few places on Earth entirely free from Covid-19 could become one of the first countries vaccinated against the disease.

The Republi ceived its first shipment of the vaccin Saturday. Vaccinations started the r

The first ship ministered in two shots, 28 days apa will be among the first to receive the vaccine, according to the ministry's statement.

To date, Palau has not recorded a single coronavirus case or virus-related death, according to the World Health Organization.

In January, as the virus began to spread across Asia and the Pacific, Palau was among the first to implement stricter border controls. Its borders were entirely locked by March, and it began testing citizens for the virus by April. These measures were key to keeping Covid-19 out, Palau's ambassador to the UN said in May.

An independent nation in free association with Washington, Palau has access to the United States' mass Covid-19 vaccination program, known as Operation Warp Speed (OWS).

The archipelago covers an area of just 177 square miles (459 square kilometers) -- about a sixth of the size of Rhode Island, the smallest US state. That small size puts Palau in prime position to be among the first nations to be inoculated against Covid-19, according to the country's incident commander of the Ministry of Health, Ritter Udui.

"We are lucky to be in a position where we have access to vaccines through OWS, and our small size makes it easier for us to roll out the program," Udui said. "It's not compulsory to receive the vaccine, so our goal is to vaccinate about 80% of the population. We hope to achieve herd immunity (through the vaccination program)."

Palau initially planned to have the vaccinations completed by May, but Udui said this deadline would "probably be extended" due to a slowdown in distribution from the US.

Image from cnn.com

Objective

Create summaries from the CNN stories dataset

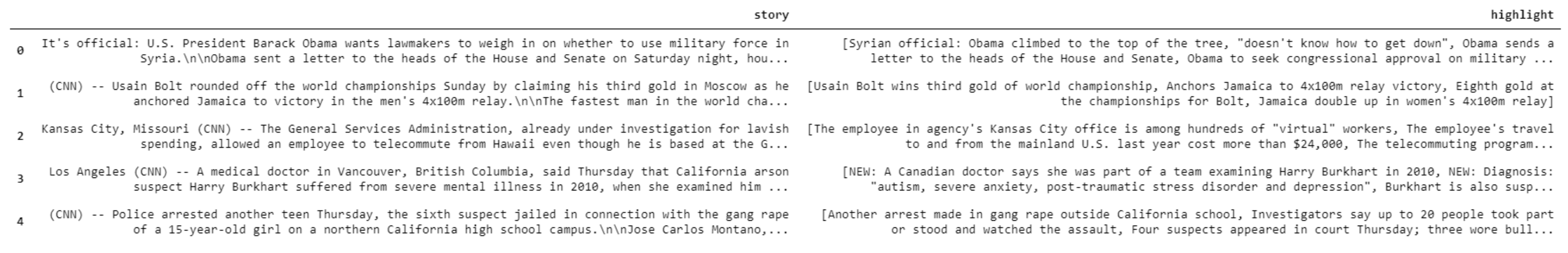
Original dataset

The dataset was taken from the DeepMind Q&A Dataset and can be found at this link. The dataset presented some issues that had to be solved before applying our models:

1) Each story presented more than one summary

2) The text was taken from the CNN website, so it contained abbreviations, html symbols, contractions, non alphabetic characters, etc… and was not suitable as input for our Summarization model

3) Only a subset of stories had to be selected for computational costraints



	story	highlight
0	It's official: U.S. President Barack Obama wants lawmakers to weigh in on whether to use military force in Syria.\n\nObama sent a letter to the heads of the House and Senate on Saturday night, hou...	[Syrian official: Obama climbed to the top of the tree, "doesn't know how to get down", Obama sends a letter to the heads of the House and Senate, Obama to seek congressional approval on military ...
1	(CNN) -- Usain Bolt rounded off the world championships Sunday by claiming his third gold in Moscow as he anchored Jamaica to victory in the men's 4x100m relay.\n\nThe fastest man in the world cha...	[Usain Bolt wins third gold of world championship, Anchors Jamaica to 4x100m relay victory, Eighth gold at the championships for Bolt, Jamaica double up in women's 4x100m relay]
2	Kansas City, Missouri (CNN) -- The General Services Administration, already under investigation for lavish spending, allowed an employee to telecommute from Hawaii even though he is based at the G...	[The employee in agency's Kansas City office is among hundreds of "virtual" workers, The employee's travel to and from the mainland U.S. last year cost more than $24,000, The telecommuting program...
3	Los Angeles (CNN) -- A medical doctor in Vancouver, British Columbia, said Thursday that California arson suspect Harry Burkhart suffered from severe mental illness in 2010, when she examined him ...	[NEW: A Canadian doctor says she was part of a team examining Harry Burkhart in 2010, NEW: Diagnosis: "autism, severe anxiety, post-traumatic stress disorder and depression", Burkhart is also susp...
4	(CNN) -- Police arrested another teen Thursday, the sixth suspect jailed in connection with the gang rape of a 15-year-old girl on a northern California high school campus.\n\nJose Carlos Montano,...	[Another arrest made in gang rape outside California school, Investigators say up to 20 people took part or stood and watched the assault, Four suspects appeared in court Thursday; three wore bull...

Solutions

The solutions that had been applied were:

1. Each row now contained one story and one summary. This meant repeating the stories where necessary.
2. Normalization
    1. Non-ascii characters removal
    2. All to lowercase
    3. Punctuation removal
    4. Numbers removal
    5. Contractions mapping
    6. Stop-words removal
3. Lemmatization
4. Tokenization
5. 10,000 stories are selected

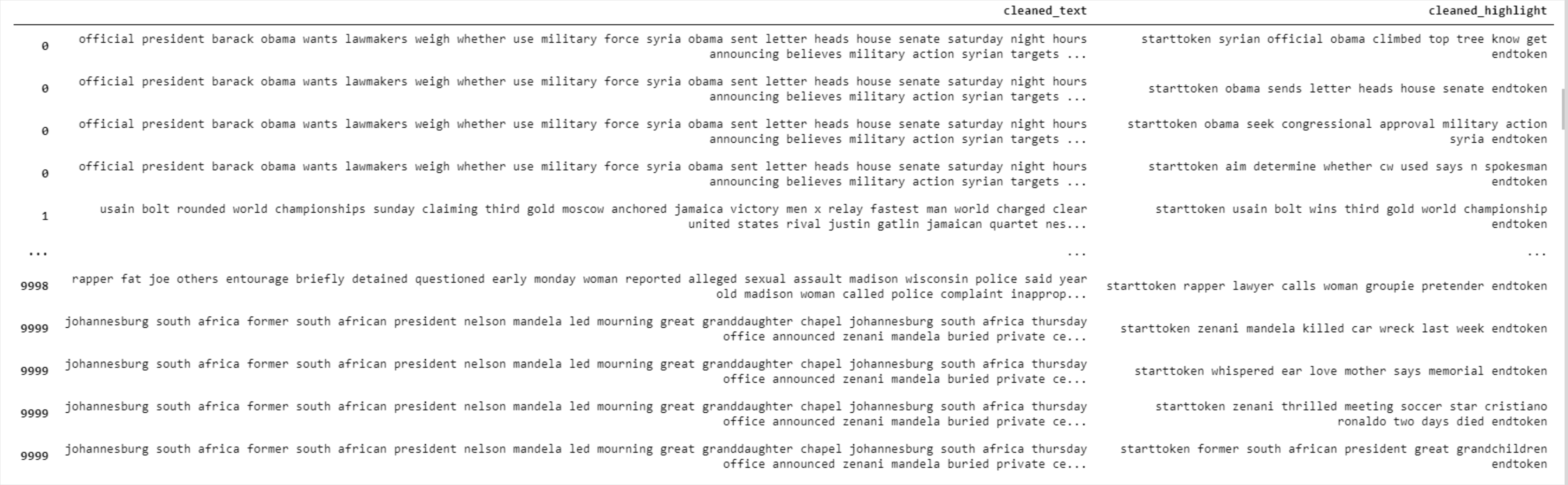
Additional steps

Additional preprocessing steps required by the nature of the problem and of the model used are applied, like:

- Adding start and stop tokens at the beginning and at the end of each summary

Resulting dataset

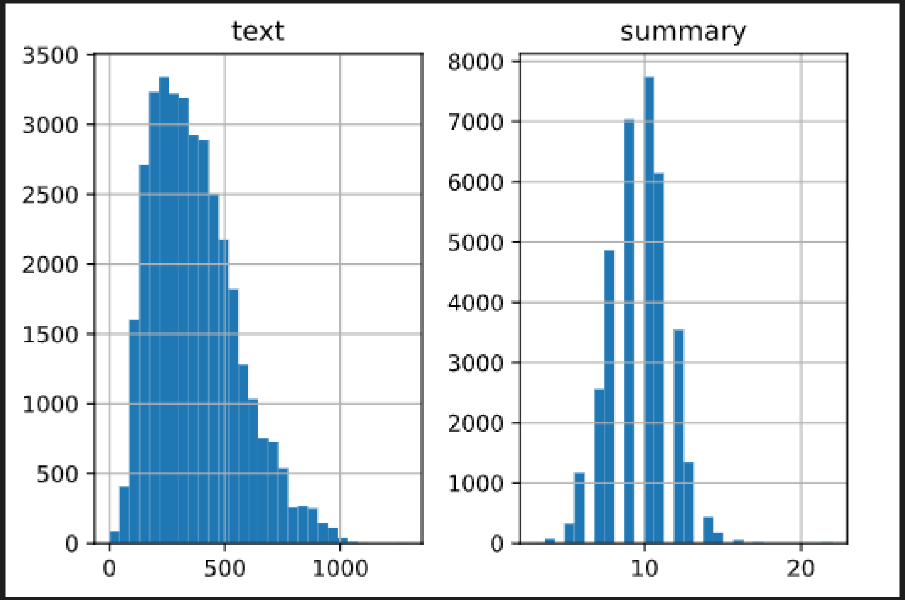
Here is the resulting dataset after all the preprocessing steps:



	cleaned_text	cleaned_highlight
0	official president barack obama wants lawmakers weigh whether use military force syria obama sent letter heads house senate saturday night hours announcing believes military action syrian targets ...	starttoken syrian official obama climbed top tree know get endtoken
0	official president barack obama wants lawmakers weigh whether use military force syria obama sent letter heads house senate saturday night hours announcing believes military action syrian targets ...	starttoken obama sends letter heads house senate endtoken
0	official president barack obama wants lawmakers weigh whether use military force syria obama sent letter heads house senate saturday night hours announcing believes military action syrian targets ...	starttoken obama seek congressional approval military action syria endtoken
0	official president barack obama wants lawmakers weigh whether use military force syria obama sent letter heads house senate saturday night hours announcing believes military action syrian targets ...	starttoken aim determine whether cw used says n spokesman endtoken
1	usain bolt rounded world championships sunday claiming third gold moscow anchored jamaica victory men x relay fastest man world charged clear united states rival justin gatlin jamaican quartet nes...	starttoken usain bolt wins third gold world championship endtoken
...	...	...
9998	rapper fat joe others entourage briefly detained questioned early monday woman reported alleged sexual assault madison wisconsin police said year old madison woman called police complaint inapprop...	starttoken rapper lawyer calls woman groupie pretender endtoken
9999	johannesburg south africa former south african president nelson mandela led mourning great granddaughter chapel johannesburg south africa thursday office announced zenani mandela buried private ce...	starttoken zenani mandela killed car wreck last week endtoken
9999	johannesburg south africa former south african president nelson mandela led mourning great granddaughter chapel johannesburg south africa thursday office announced zenani mandela buried private ce...	starttoken whispered ear love mother says memorial endtoken
9999	johannesburg south africa former south african president nelson mandela led mourning great granddaughter chapel johannesburg south africa thursday office announced zenani mandela buried private ce...	starttoken zenani thrilled meeting soccer star cristiano ronaldo two days died endtoken
9999	johannesburg south africa former south african president nelson mandela led mourning great granddaughter chapel johannesburg south africa thursday office announced zenani mandela buried private ce...	starttoken former south african president great grandchildren endtoken

Methodological approach and evaluation

First, we selected a max length of text (300) to take and a max summary length (12) based on the distribution of the data.



Then we implemented an Abstractive Summarization model.

Structure:

- Encoder
- Decoder
- Attention Layer

The first model

The model used is structurally simple.

The encoder is composed of an Embedding layer followed by a single LSTM layer.

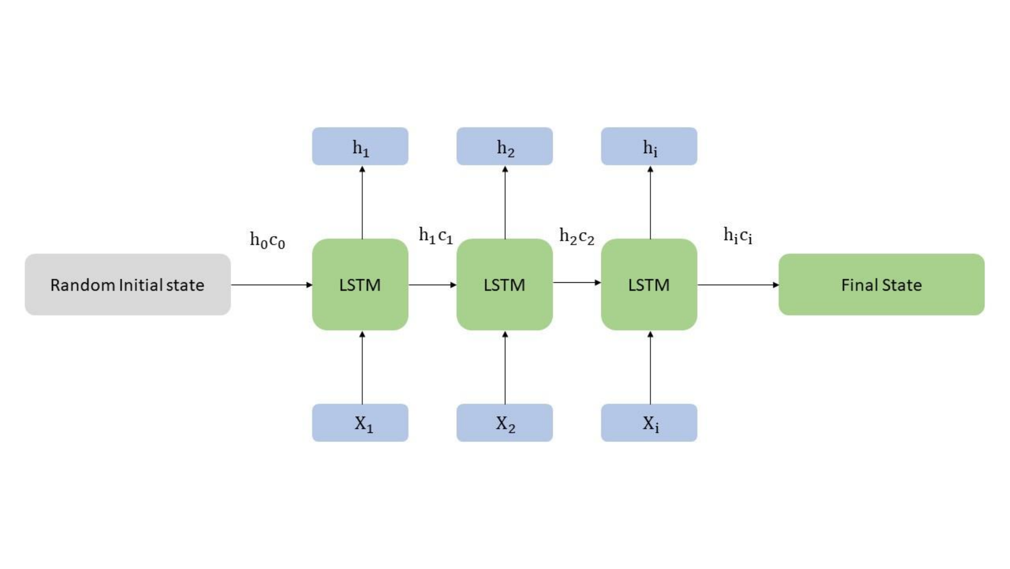
The decoder is composed similarly to the encoder, with one Embedding and one LSTM layer.

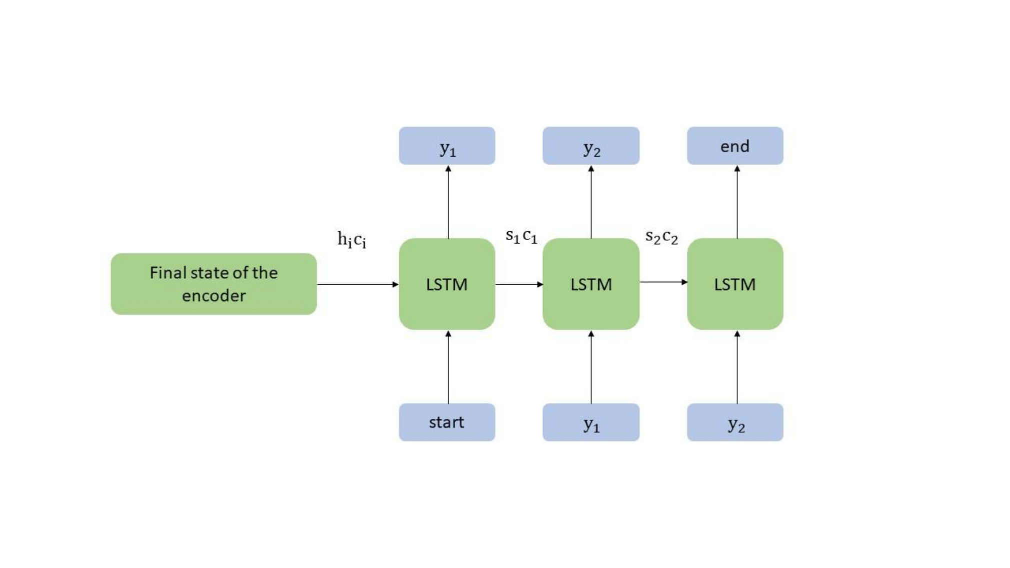
Between the two, an Attention layer is placed.

Then, a Time-distributed Dense layer is added at the end.

Although the relatively simple structure, this model contained more than 10 Million parameters that had to be trained locally on our machines.

This explains why only 20 epochs has been used to train the model. Each epoch took approximately 1 hour.





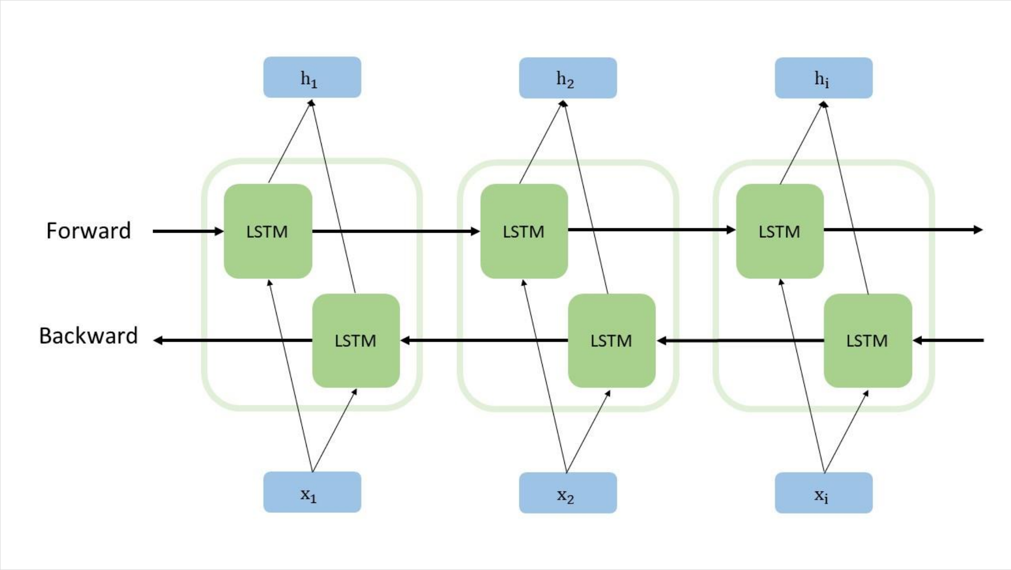
The second model

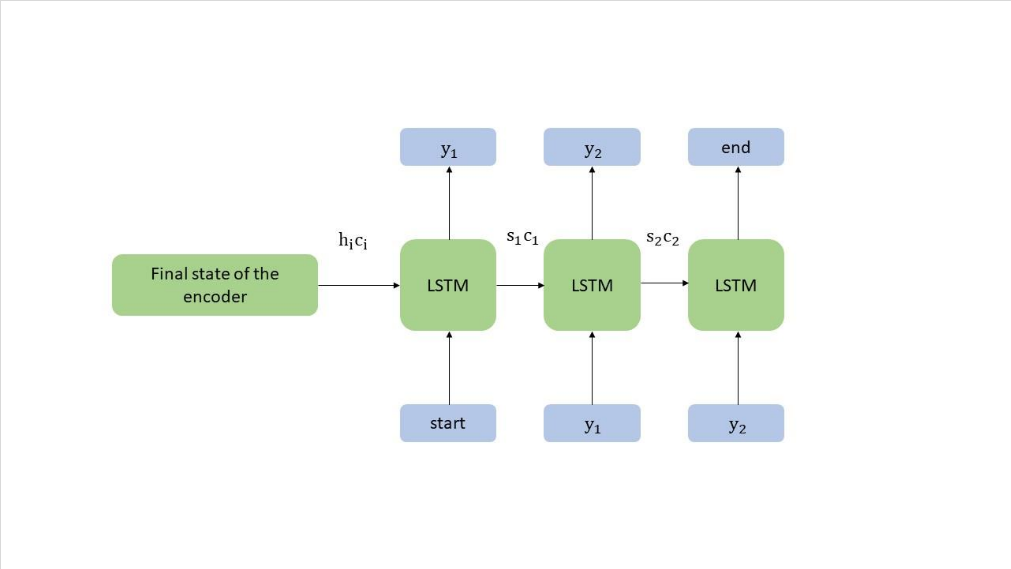
The second model improved from the first model by implementing a bidirectional LSTM encoder,

while maintaining a unidirectional decoder.

The forward and reverse hidden states of the bidirectional encoder are then concatenated two by two and then fed as initial states to the unidirectional decoder.

This should give the model better understanding of the patterns in the text.

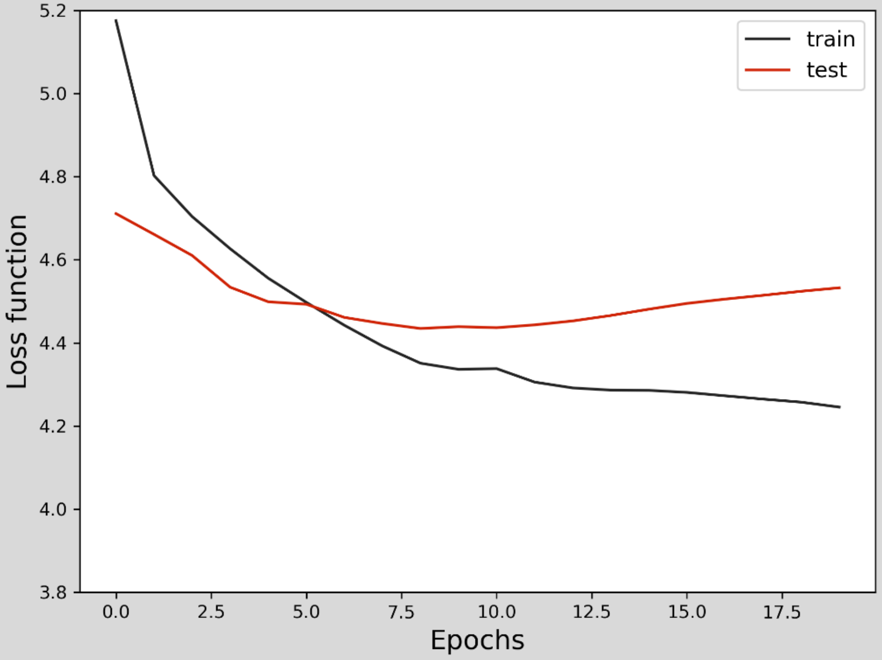




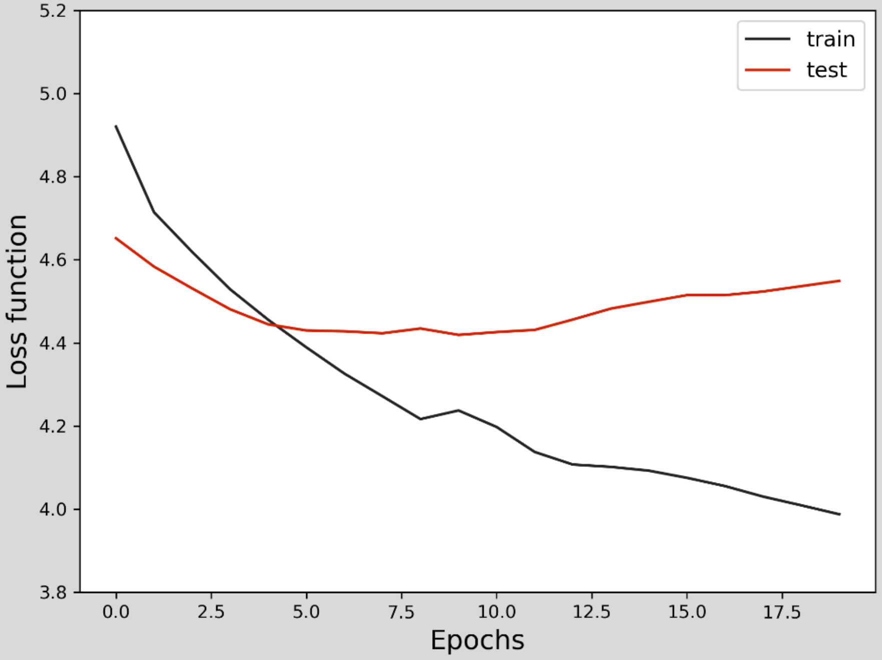
Results

The following graphs represent the loss function of the two models

Monodirectional model



Bidirectional model



Results

Monodirectional model

ROUGE-1: 7.56%

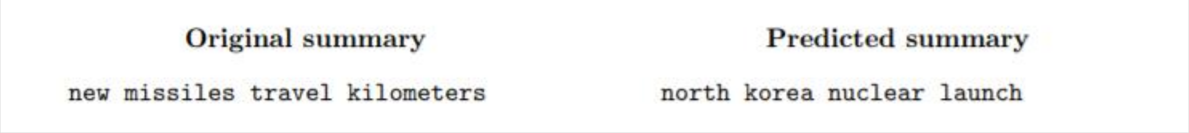
ROUGE-2: 0.80%

Bidirectional model

ROUGE-1: 6.62%

ROUGE-2: 0.66%

Example



Original summary	Predicted summary
new missiles travel kilometers	north korea nuclear launch

The two pre-trained models

BART

This model uses a standard seq2seq/machine translation architecture with a bidirectional encoder (like BERT) and a left-to-right decoder (like GPT).

This model partially reflects the structure of the model built by us.

T5

T5 is a Google pre-trained model, which propose reframing all NLP tasks into a unified text-to-text format where the input and output are always text strings, in contrast to BERT-style models that can only output either a class label or a span of the input.

Results

BART model

ROUGE-1: 17.46%

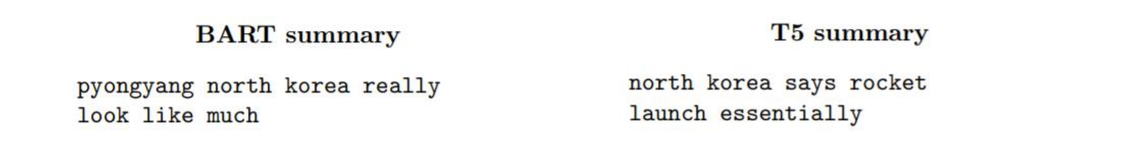
ROUGE-2: 7.23%

T5 model

ROUGE-1: 20.51%

ROUGE-2: 8.81%

Example



BART summary	T5 summary
pyongyang north korea really look like much	north korea says rocket launch essentially

Conclusions

It is clear how these Seq2Seq models need more training time and far more computational power than what is available to a normal developer/consumer.

However, what stood out the most is that they were capable to produce a few summaries that impressed, given the limited resources that we had to deal with.

In particular, the model that made use of the bidirectional LSTM, although presenting lower ROUGE scores, generated more complex and fluent outputs, and confirms its greater capacity in recognizing patterns in text.

Thank You

Christian Uccheddu - 800428
Federico De Servi - 812166

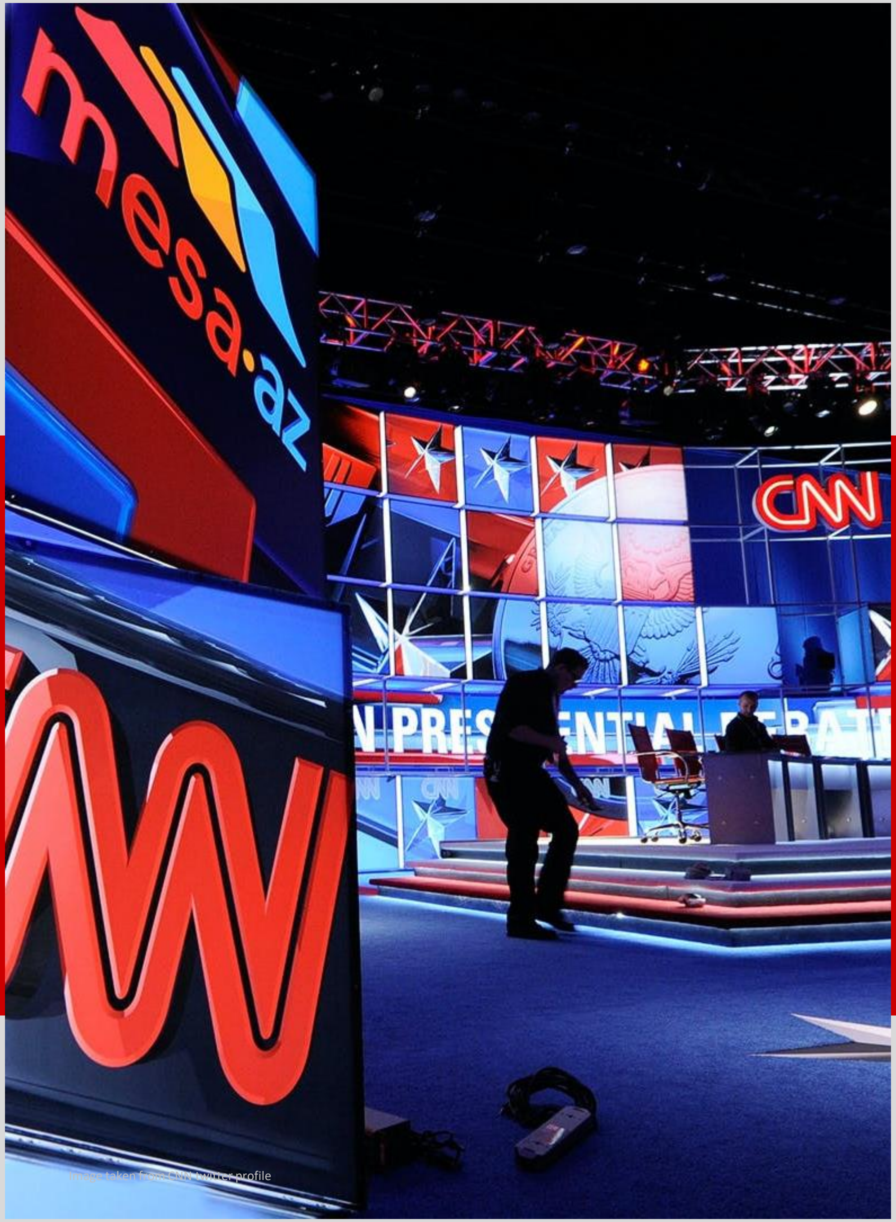


Image taken from CNN twitter profile