Text Summarization using Natural Language Processing(NLP)

Mayank Jain 2019A7PS0141P
Aryan Devrani 2019A8PS0408P

What is NLP?

- A field of Artificial Intelligence (AI) that makes human language intelligible to machines.
- Combines the power of linguistics and computer science to study the rules and structure of language, and create intelligent systems capable of :

  1. Understanding

  2. Analyzing, and

  3. Extracting meaning from text and speech.

What is NLP used for ?

To understand the structure and meaning of human language by analyzing aspects like:

- Syntax
- Semantics
- Pragmatics
- Morphology

Applications:

- Performing large-scale analysis of unstructured text data
- Automating information sort/route processes in real-time
- Tailored NLP tools (Sarcasm, Misused words)

Dataset

- The CNN / DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail.
- The current version supports both extractive and abstractive summarization
- Version 3.0.0:

Dataset Split	Number of Instances in Split
Train	287,113
Validation	13,368
Test	11,490

Preprocessing

1. Cleaning-Removing punctuation and characters
2. Tokenizing - Convert from string to list
3. Stemming - Removing ending phrases like -ing, -ily etc.
4. Lemmatization - Convert the word into root word

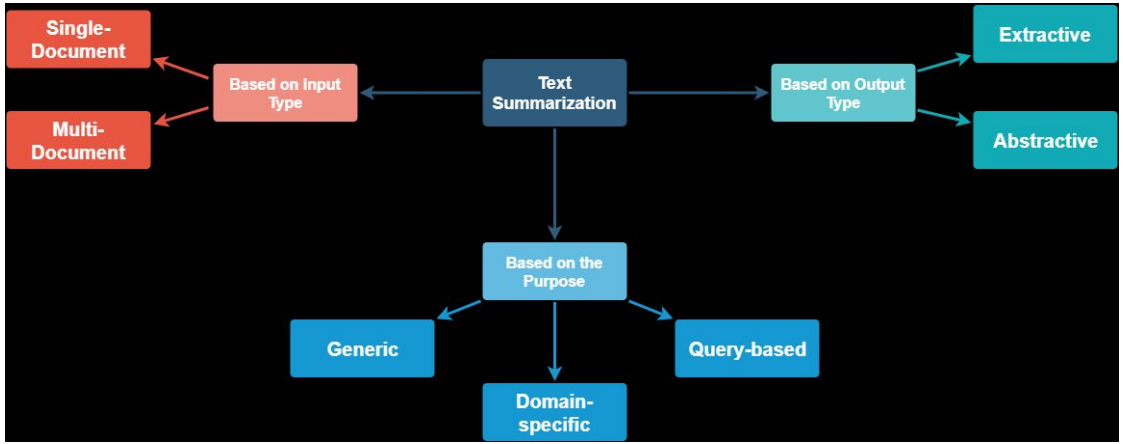
Metric for comparing results

ROUGE score (Recall-Oriented Understudy for Gisting Evaluation)
It measures the overlap between the system summary and the reference summary in terms of consecutive tokens a.k.a. n-grams. It uses the F1-score for capturing more information.
A large overlap of n-grams results in a high ROUGE score and vice-versa.

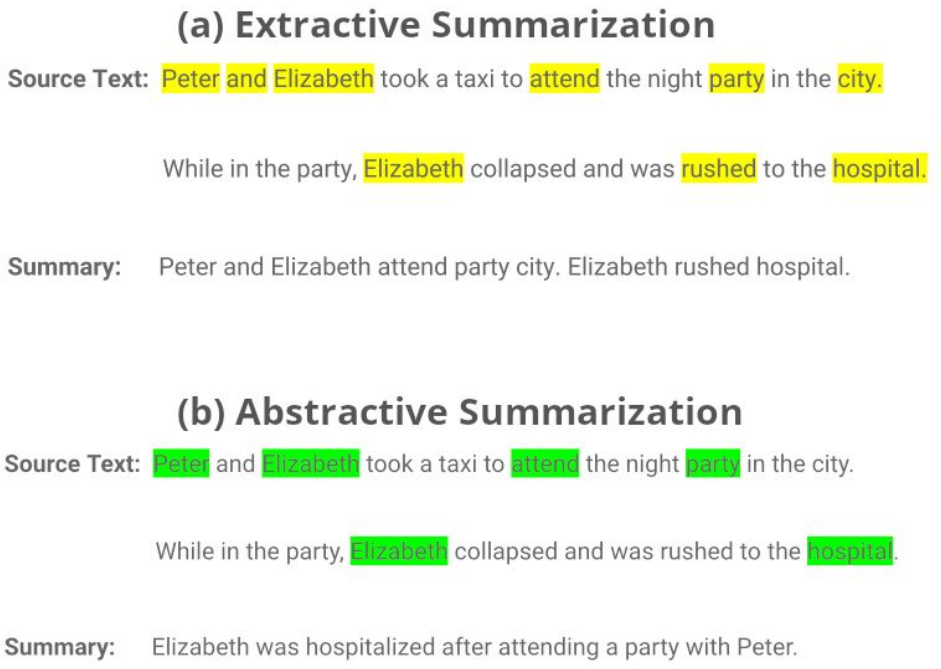
We use the average of 3 ROUGE scores -
* ROUGE-1 refers to overlap of unigrams between the two summaries.
* ROUGE-2 refers to the overlap of bigrams between the two summaries.
* Rouge-L measures longest matching sequence of words using Longest Common Subsequence.

Text Summarization in NLP

- The process of creating a short, coherent, and fluent summary of a longer text document outlining the text's major points.



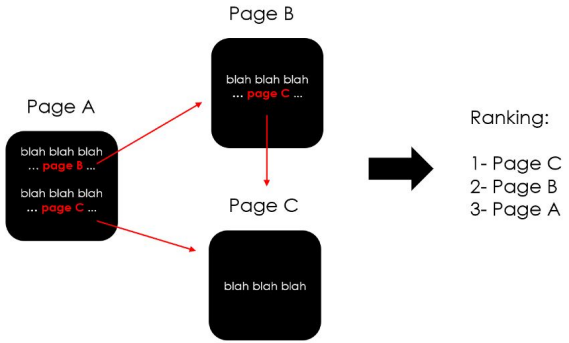
Extractive and Abstractive Summarization



TextRank (Extractive Summarization)

- Graph-based ranking model, based on Google's **PageRank** algorithm, that finds the most relevant sentences in a text. Used by Google search engine to sort web pages in 1998.

1. Segment document into paragraphs and sentences
2. Parse each sentence (NLP) : Part of Speech (PoS), root word (lemma), meaning
3. Construct a graph with nouns, adjectives and verbs as the vertices.
    - Links based on skip-grams (predict context word for a given target word)
    - Links based on repeated instances of a lemma (root word)
    - Links inferred from knowledge graph
4. Eigenvalue Centrality (**PageRank)**



TextRank Results

```
[ ]  # Compare y_test and predicted
     match = display_string_matching(dtf_test["y"][i], predicted[i], both=True, sentences=False,
                                     titles=["Real Summary", "Predicted Summary"])

     from IPython.core.display import display, HTML
     display(HTML(match))
```

Real Summary
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 bulletproof vests .

Predicted Summary
(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.

```
     # Explainability
     match = display_string_matching(dtf_test["text"][i], predicted[i], both=True, sentences=True,
                                     titles=["Full Text", "Predicted Summary"])

     from IPython.core.display import display, HTML
     display(HTML(match))
```

Full Text
(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. Jose Carlos Montano, 18, was arrested on charges of felony rape, rape in concert with force, and penetration with a foreign object, said Richmond Police Lt. Mark Gagan. Montano was arrested Thursday evening in San Pablo, California, a small town about two miles from the city of Richmond, where the crime took place. Montano, who was held in lieu of $1.3 million bail, is accused of taking part in what police said was a 2½-hour assault on the Richmond High School campus. Police said as many as 10 people were involved in the rape in a dimly lit back alley at the school, while another 10 people watched without calling 911. The victim was taken to the hospital in critical condition, but was released Wednesday. Four other teenage suspects were arraigned Thursday on charges connected to the rape. Cody Ray Smith, described by the court as older than 14, pleaded not guilty to charges of rape with a foreign object and rape by force. Two other juveniles, Ari Abdallah Morales and Marcelles James Peter, appeared with Smith at the Contra Costa County Superior Court, but did not enter a plea. The court described Morales as younger than 16, and did not give an age for Peter. All three juveniles, who wore bulletproof vests at the hearing, were charged as adults. A fourth person, Manuel Ortega, 19, appeared separately without an attorney and did not enter a plea. He did not wear a protective vest. Another person, Salvador Rodriguez, 21, was arrested Tuesday night, but he was not in court Thursday.

Predicted Summary
(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.

Seq2Seq (Abstractive)

These models convert take a sequence from one domain (text vocabulary) and output a new sequence in another domain(summary vocabulary).

These models have the following characteristics -

1. Sequences as corpus
2. Word embedding Mechanism - Words from vocabulary are mapped to vectors of real numbers which are calculated from the probability distribution for each word appearing before or after another.
3. 2 models - One for training, other for inference/predictions. The prediction model leverages the encoder from the trained model.
4. Encoder-Decoder structure - The encoder processes the input sequence and returns its own internal states that serve as the context for the decode, which predicts the next word of the target sequence, given the previous ones.

Seq2Seq

Methodology-

1. Divide Text corpus into Sequences -
   The texts are padded into sequences with the same length to get a feature matrix.
2. Create Embeddings -
3. Training the Encoder Decoder model -
   Inputs - *X* (the text sequences) plus the *y* (summary sequences), we are just gonna hide the last word of the summaries
   Target – *y* (summary sequences) without the *start* token.

4. Prediction Decoder -
Inputs - <START> token & outputs of the encoder and its states
Output - New States as well as a probability distribution over the vocabulary
from which (the word with the highest probability will be the prediction).
The Decoder uses the generated word and the new states to predict the new word and new states. This iteration shall go on until the model finally predicts the *end* token or the predicted summary reaches its maximum length.

Seq2Seq Results

Real Summary
usain bolt win third gold world championship anchor jamaica 4x100m relay victory eighth gold championship bolt jamaica double woman 4x100m relay

Predicted Summary
Earlier, Jamaica's women underlined their dominance in the sprint events by winning the 4x100m relay gold, anchored by Shelly-Ann Fraser-Pryce, who like Bolt was completing a triple.

```
[ ]  # Explainability
     match = display_string_matching(dtf_test["text_clean"][i], predicted[i], both=True, sentences=False,
                                     titles=["Full Text", "Predicted Summary"])

     from IPython.core.display import display, HTML
     display(HTML(match))
```

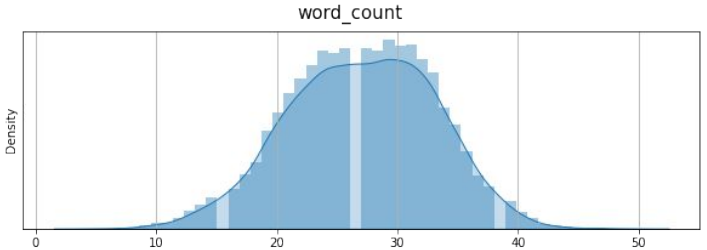
Full Text
usain bolt rounded world championship sunday claiming third gold moscow anchored jamaica victory men 4x100m relay fastest man world charged clear united state rival justin gatlin jamaican quartet nesta carter kemar baileycole nickel ashmeade bolt 3736 second finished second 3756 second canada taking bronze britain disqualified faulty handover 26yearold bolt collected eight gold medal world championship equaling record held american trio carl lewis michael johnson allyson felix mention small matter six olympic title relay triumph followed individual success 100 200 meter russian capital proud ill continue work dominate long possible bolt previously expressed intention carry 2016 rio olympics victory never seriously doubt got baton safely hand ashmeade gatlin united state third leg runner rakieem salaam problem gatlin strayed lane struggled get full control baton never able get term bolt earlier jamaica woman underlined dominance sprint event winning 4x100m relay gold anchored shellyann fraserpryce like bolt completing triple quartet recorded championship record 4129 second well clear france crossed line second place 4273 second defending champion united state initially back bronze medal position losing time second handover alexandria anderson english gardner promoted silver france subsequently disqualified illegal handover british quartet initially fourth promoted bronze eluded men team fraserpryce like bolt aged 26 became first woman achieve three gold 100200 relay final action last day championship france teddy tamgho became third man leap 18m triple jump exceeding mark four centimeter take gold germany christina obergfoll finally took gold global level woman javelin five previous silver kenya asbel kiprop easily tactical men 1500m final kiprops compatriot eunice jepkoech sum surprise winner woman 800m bolt final dash golden glory brought eightday championship rousing finale host topped medal table united state criticism poor attendance luzhniki stadium concern pole vault gold medalist yelena isinbayeva made controversial remark support russia law make propagandizing nontraditional sexual relation among minor criminal offense later attempted clarify comment renewed call gay right group boycott 2014 winter game sochi next major sport event russia

Predicted Summary
Earlier, Jamaica's women underlined their dominance in the sprint events by winning the 4x100m relay gold, anchored by Shelly-Ann Fraser-Pryce, who like Bolt was completing a triple.

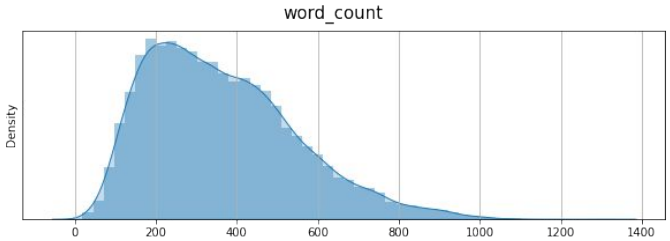
Feature Engineering

Since we need to transform text into sequences of words, we need to decide -

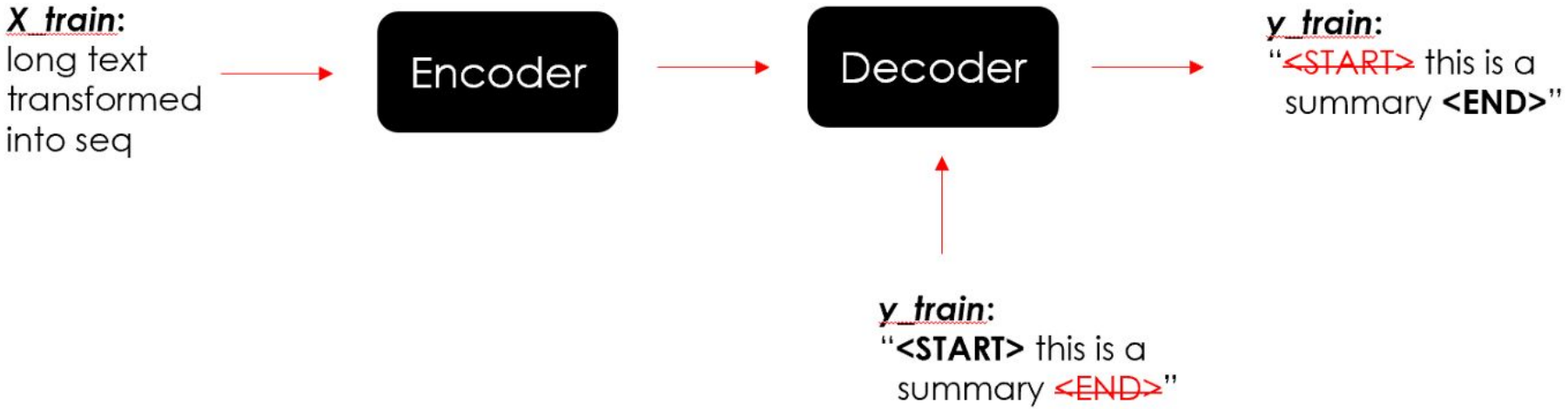
- Optimal sequence size
- Number of words our model should remember.



Text



Summary



Training Model

Inference Model

Transformer Model in NLP

- Original transformer based on encoder-decoder architecture.
- Encoder learns a high-dimensional representation of the input, which is mapped to the output by the decoder.
- Nowadays, transformer models are pre-trained on a large text corpus (e.g Wikipedia)
- This ensures the ability of the model to understand language, for text classification or summarization.

BART (Bidirectional Auto-Regressive Transformer)

- A pre-trained model by Facebook (Abstractive Summarization)
- Is a Sequence-to-sequence model trained as a denoising autoencoder.
- What counts as noise for text data?
- Noising schemes:
    1. Token Masking
    2. Token Deletion
    3. Text Infilling
    4. Sentence Permutation
    5. Document Rotation

BART Results

```
[ ]
from IPython.core.display import display, HTML
display(HTML(match))
```

Real Summary
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 bulletproof vests .

Predicted Summary
Jose Carlos Montano, 18, was arrested Thursday evening in San Pablo, California. Montano is accused of taking part in what police said was a 2½-hour assault on a

```
# Explainability
match = display_string_matching(dtf_test["text"][i], predicted[i], both=True, sentences=True,
                                titles=["Full Text", "Predicted Summary"])

from IPython.core.display import display, HTML
display(HTML(match))
```

Full Text
(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. Jose Carlos Montano, 18, was arrested on charges of felony rape, rape in concert with force, and penetration with a foreign object, said Richmond Police Lt. Mark Gagan. Montano was arrested Thursday evening in San Pablo, California , a small town about two miles from the city of Richmond, where the crime took place. Montano, who was held in lieu of $1.3 million bail, is accused of taking part in what police said was a 2½-hour assault on the Richmond High School campus. Police said as many as 10 people were involved in the rape in a dimly lit back alley at the school, while another 10 people watched without calling 911. The victim was taken to the hospital in critical condition, but was released Wednesday. Four other teenage suspects were arraigned Thursday on charges connected to the rape. Cody Ray Smith, described by the court as older than 14, pleaded not guilty to charges of rape with a foreign object and rape by force. Two other juveniles, Ari Abdallah Morales and Marcelles James Peter, appeared with Smith at the Contra Costa County Superior Court, but did not enter a plea. The court described Morales as younger than 16, and did not give an age for Peter. All three juveniles, who wore bulletproof vests at the hearing, were charged as adults. A fourth person, Manuel Ortega, 19, appeared separately without an attorney and did not enter a plea. He did not wear a protective vest. Another person, Salvador Rodriguez, 21, was arrested Tuesday night, but he was not in court Thursday.

Predicted Summary
Jose Carlos Montano, 18, was arrested Thursday evening in San Pablo, California . Montano is accused of taking part in what police said was a 2½-hour assault on a

Results

Model	Final ROUGE Score
TextRank	0.166
Seq2Seq	0.082
BART	0.286

- A higher ROUGE Score essentially implies a better overlap between the system and reference summaries.
- BART, a pre-trained transformer based model, performs the best compared to TextRank and Seq2Seq