TEXT SUMMARIZATION SYSTEM

ANLY-580 Natural Language Processing

Group Members: Hanshen Jing | Peijin Li | Zihang Weng | Zixuan Wang

TABLE OF CONTENTS

01

TEXT SUMMARIZATION

- INTRODUCTION
- TYPES OF SUMMARIZATION
- PROS & CONS OF ABSTRACTIVE SUMMARIZATION

02

MODELS

- SEQ2SEQ MODEL: BERT
- MODEL EVALUATION METRICS: ROUGE

03

OUR MODEL TRAINING PROCESS

- DATASET
- DATA PRE-PROCESSING & EDA
- MODEL & MODEL EVALUATION
- DEMO

TEXT SUMMARIZATION

01

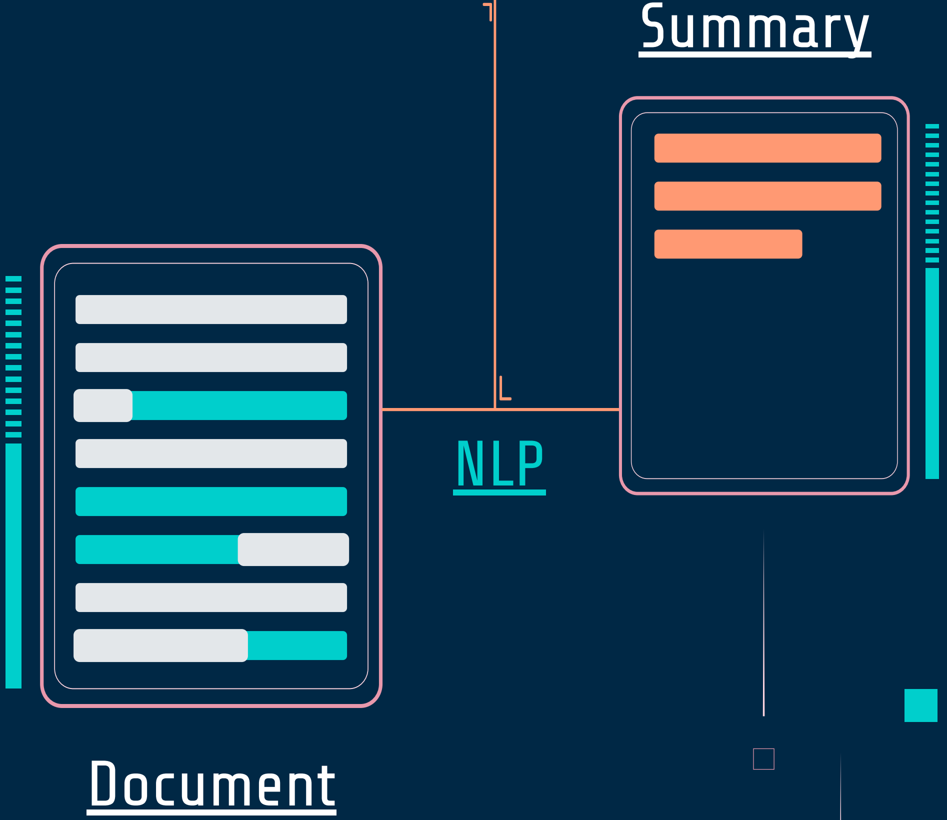
- INTRODUCTION
- TYPES OF SUMMARIZATION
- PROS & CONS OF ABSTRACTIVE SUMMARIZATION

TEXT SUMMARIZATION

GOAL: To produce a shorter version of a source, which conveys the essence of the document, helps in finding relevant information quickly

USE CASES

- Email overload
- Science and R&D
- Books and literature
- ...



TYPES OF SUMMARIZATION

There are two types summaries

EXTRACTIVE SUMMARIZATION

- Created by reusing portions (words, sentences, etc.) of the input text document.
- The system extracts text from the entire collection, without modifying the text document.

ABSTRACTIVE SUMMARIZATION

- Generates own summary over input text without using same words or sentence in the input text.
- Requires deep understanding and reasoning over the text.

ABSTRACTIVE SUMMARIZATION

PROS & CONS

- Abstractive summarization leverages contextual learning to generate powerful summaries. These summaries are more human-readable, making them easier to consume.

- The abstractive approach involves summarization based on deep learning. It is more computationally expensive.

MODELS

- SEQ2SEQ MODEL: BART
- MODEL EVALUATION METRICS: ROUGE

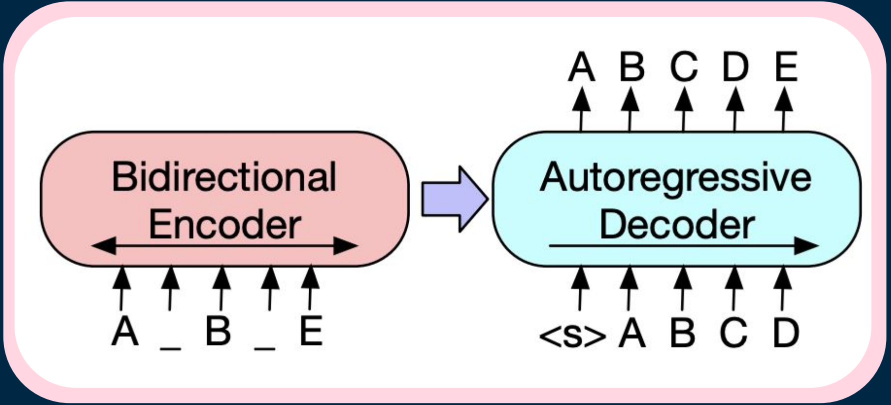
02

Seq2seq PRE-TRAINED MODELS: BART

BART is a denoising autoencoder for pretraining sequence-to-sequence models.

It is trained by

- corrupting text with an arbitrary noising function
- learning a model to reconstruct the original text



It uses a standard Transformer-based neural machine translation architecture. It uses a standard seq2seq/NMT architecture with a bidirectional encoder (like BERT) and a left-to-right decoder (like GPT). This means the encoder's attention mask is fully visible, like BERT, and the decoder's attention mask is causal, like GPT2.

EVALUATING SUMMARIES: ROUGE

$$\text{ROUGE}_{\text{RECALL}} = \frac{\text{NUMBER OF OVERLAPPING WORDS}}{\text{TOTAL WORDS IN REFERENCE SUMMARY}}$$

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation

A set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing

The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation.

EVALUATING SUMMARIES: ROUGE (continued)

ROUGE-N

Overlap of n-grams between the system and reference summaries. E.G. ROUGE-1, ROUGE-2

ROUGE-L

Longest Common Subsequence (LCS) based statistics.

ROUGE-W

Weighted LCS-based statistics that favors consecutive LCSes.

ROUGE-S

Skip-bigram based co-occurrence statistics. Skip-bigram is any pair of words in their sentence order.

ROUGE-SU

Skip-bigram plus unigram-based co-occurrence statistics.

OUR MODEL TRAINING PROCESS

03

- DATASET
- DATA PRE-PROCESSING & EDA
- MODEL & MODEL EVALUATION
- DEMO

DATASET

CNN/DailyMail

News articles of CNN and Daily Mail and corresponding summaries (the highlight of the article as written by the article author).

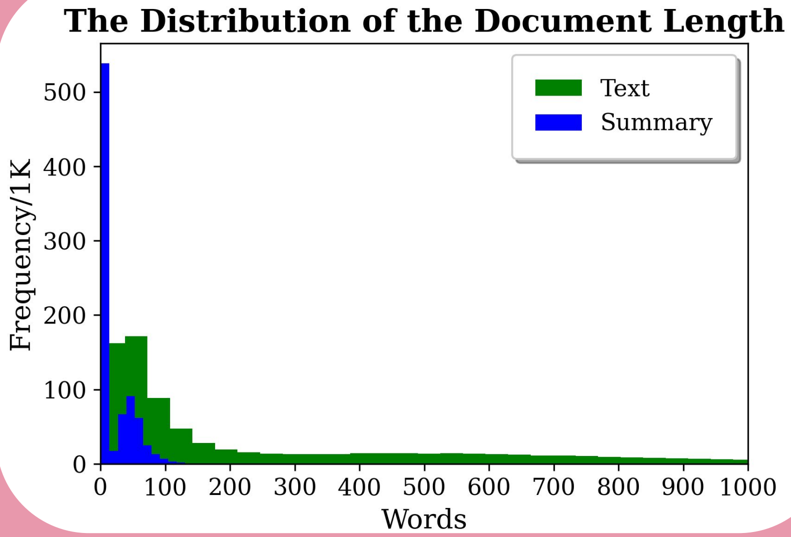
Size: >300k records

Amazon Fine Food Reviews

Reviews of fine foods from Amazon, including product and user information, ratings, and a plain text review.

Size: >500k records

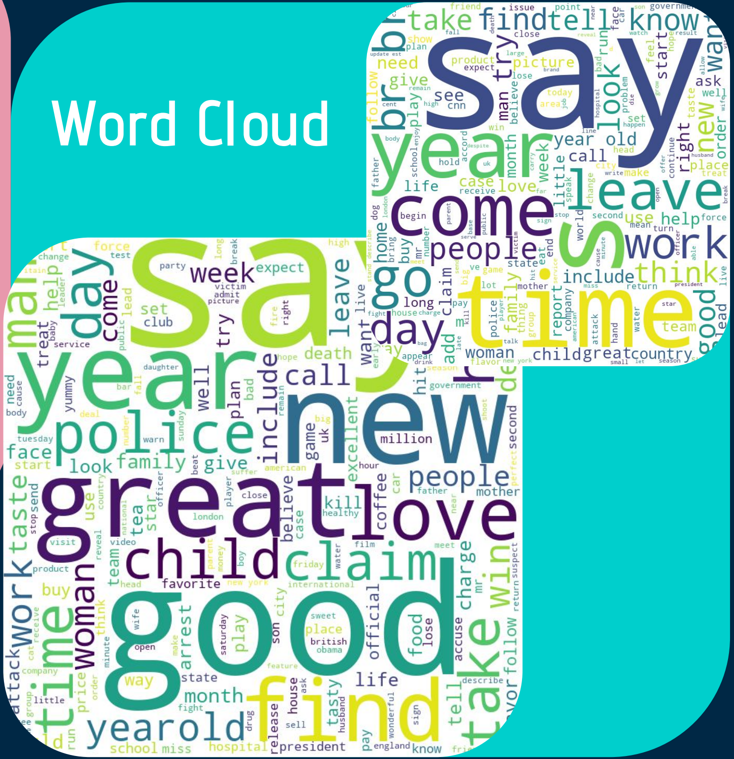
EXPLORATORY DATA ANALYSIS



DATA SELECTION

- ORIGINAL TEXT < 200 WORDS
- SUMMARY < 60 WORDS

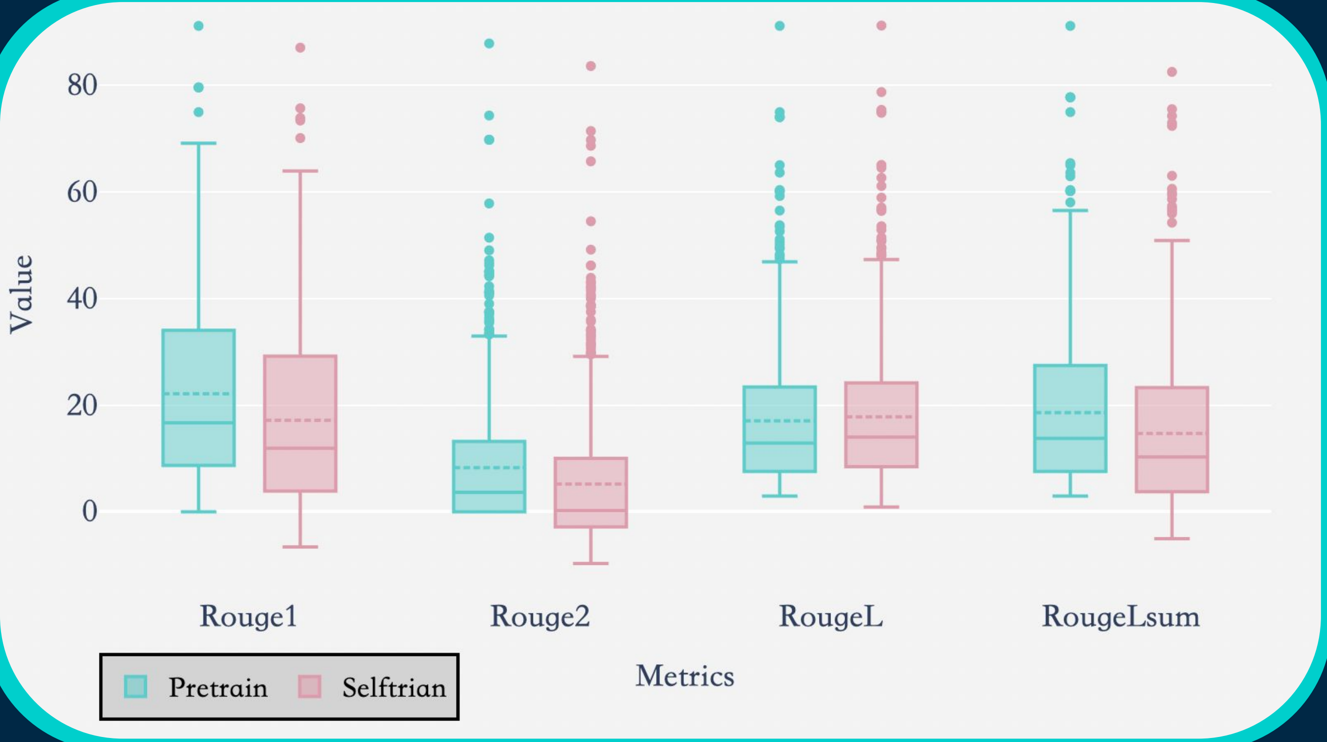
Word Cloud



TWO BART-BASED MODELS

-	XSum	For Shorter Summarization
-	CNN/Dailymail	For Longer Summarization

THE EVALUATION OF TWO MODELS



DEMO

NPR News:
https://www.npr.org/2022/11/30/1139742011/jeffries-poised-to-make-history-as-first-black-person-to-lead-congressional-part

The New York Times:
https://www.nytimes.com/2022/11/29/us/politics/biden-rail-strike.html

Best Restaurant in DC by The Washington Post:
https://www.washingtonpost.com/food/interactive/2022/best-restaurants-in-dc-2022/

THANKS

Do you have any questions?