CAS2105 Homework 6: Evaluating Neural Summarization Against a Lead-1 Baseline Using ROUGE Metrics

Birken Michael Silitch (2024149080)

1 Introduction

In this project, I chose to work on the task of **text summarization**. The goal of this task is simple: given a long news article, the system should generate a much shorter summary that still captures the main idea. Summarization is interesting because it is something people do naturally all the time—for example, when we tell a friend what happened in an article we read, or when we shorten a long message into a few key points. Building a computer system that can do this automatically requires understanding language, identifying important information, and ignoring irrelevant details.

This task is also a good fit for an introductory AI pipeline project because it can be approached at multiple levels of complexity. A very simple system can summarize a document just by returning the first sentence, which often contains the most important information in news writing. At the same time, modern pre-trained models can generate summaries that are more fluent and more abstract. Comparing these two approaches helps us understand not only how AI models work, but also what kinds of improvements they actually provide.

By working on summarization, I can clearly demonstrate the full workflow of an AI pipeline: defining a problem, building a baseline, using a pre-trained model, evaluating both systems, and reflecting on their strengths and weaknesses. The task is small enough to run on a single GPU, but rich enough to reveal meaningful insights about model behavior and evaluation.

2 Task Definition

Task description

The goal of my project is to perform **abstractive text summarization** on news articles. Given a full article, the system should generate a short, one–sentence summary that captures the main idea of the story. In other words, the model must read a long document and rewrite it in a much shorter form.

Motivation

Summarization is interesting because people frequently need quick, digestible versions of long texts—for example, when skimming news, reviewing documents, or catching up on missed information. Automating this process can save time and help organize information more efficiently.

Input / Output

- **Input:** A news article, usually several sentences long (100–300 words).
- **Output:** A short summary generated by the system.

Success criteria

A summarization system is considered "good" if:

- its generated summaries overlap well with human-written reference summaries (measured using **ROUGE-1** and **ROUGE-L**),
- it captures the key information of the article without adding unrelated details,
- and it produces fluent, readable English.

To quantify performance, I compare the ROUGE scores of my baseline system (Lead-1) with the scores of the pre-trained transformer model. Higher ROUGE scores indicate better summarization quality.

3 Methods

This section describes both the naïve baseline system and the improved AI pipeline used for abstractive summarization. The goal is to compare a simple heuristic approach with a modern pre-trained transformer model.

3.1 Naïve Baseline

Lead-1 Baseline

- **Method description:** I implement the classic **Lead-1** baseline, which simply returns the first sentence of the article as the summary. This method relies on a well-known property of news writing: important information is often placed at the beginning of the article. The baseline extracts the document, splits it into sentences, and outputs the first one without any additional processing.
- **Why naïve:** This baseline does not "understand" the text in any meaningful way. It does not compute importance, does not rephrase content, and does not adapt to different writing styles. It only assumes that the first sentence is always representative of the article, which is not necessarily true. The method requires almost no computation and serves as a minimal reference point for evaluating more advanced systems.
- **Likely failure modes:**
  - When the first sentence contains background context rather than the main event.
  - When key information appears later in the article.
  - When the first sentence is unusually long or contains multiple topics, reducing clarity.
  - When the article includes quotes or rhetorical openings before stating the main point.

3.2 AI Pipeline

DistilBART CNN-12-6

- **Models used:** I use the pre-trained **DistilBART CNN-12-6** model from the HuggingFace Transformers library. This is a lightweight, distilled version of BART trained specifically for news summarization tasks.
- **Pipeline stages:**
  1. **Preprocessing:** Each article is cleaned and prepared for input to the model. In practice, this step is minimal because the HuggingFace tokenizer handles lowercasing, subword splitting, and truncation.

2. **Encoding / Representation:** The document is fed into DistilBART's encoder, which produces contextualized embeddings capturing semantic information about each token.
3. **Decoding:** The decoder generates a short abstractive summary word-by-word using beam search. I set constraints such as minimum and maximum output length to avoid excessively short or truncated summaries.
4. **Post-processing:** The resulting summary string is cleaned, stripped of special tokens, and returned as the final output.

- **Design choices and justification:**
  - DistilBART is small enough to run efficiently on a single GPU (such as Vessl's environment), yet powerful enough to produce coherent abstractive summaries.
  - Using a pre-trained model avoids the need for fine-tuning while still benefiting from large-scale training on news corpora.
  - The model handles long documents better than extractive heuristics and is able to paraphrase or compress information rather than copying sentences.
  - The pipeline remains simple and modular, making it easy to evaluate or replace individual components.

4 Experiments

4.1 Datasets

`cnn_dailymail` dataset (version 3.0.0)

- **Source:** HuggingFace `cnn_dailymail` dataset (version 3.0.0). This dataset contains news articles paired with abstractive human-written summaries, making it suitable for evaluating summarization pipelines.
- **Total examples used:** 250
- **Train/Test split:** I did not train a model; only inference was performed. Therefore, I used:
  - The dataset's built-in training split to draw examples,
  - A manually created 50-sized test set for evaluating both baseline and model.
- **Preprocessing steps:**
  - Removed extremely long articles by truncating at the model's maximum input token length.
  - Applied the HuggingFace DistilBART tokenizer (handles normalization, subword splitting, integer mapping).
  - For the baseline, performed sentence splitting using a regex rule: `re.split(r'(?<=[.!?])\s+(?=[A-Z])', text)`, which separates sentences based on punctuation followed by a capital letter.
  - Used `min_length=15` and `max_length=100` in generation to avoid degenerate summaries.

4.2 Metrics

For evaluation, I use two standard summarization metrics[1]:

- **ROUGE-1** (unigram overlap): measures how many important words the system summary shares with the human summary.

- **ROUGE-L** (longest common subsequence): measures structural similarity and whether the system captures the main sequence of ideas.

These metrics align well with the task because the goal is to preserve key content from the article, even if the model paraphrases. ROUGE is widely used in summarization research and provides a fair comparison between extractive and abstractive systems.

4.3 Results

Table 1 shows the performance of the naïve Lead-1 baseline compared with the DistilBART summarization model.

Method	**ROUGE-1 (F)**	**ROUGE-L (F)**
Baseline (Lead-1)	0.2766	0.1963
AI Pipeline (DistilBART)	0.4372	0.3162

Table 1: Comparison of average ROUGE scores on the test set.

Visualizations

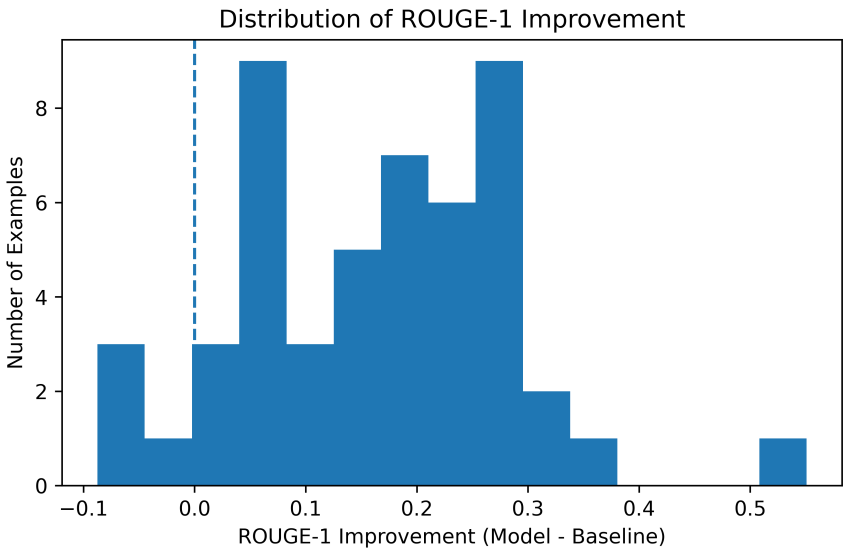

Figure 1: Distribution of ROUGE-1 Score (Model – Baseline).The histogram shows that the pretrained model outperforms the baseline on most examples. However, there are a few negative outliers where the baseline scores higher, typically when the first sentence of the article closely matches the reference summary.

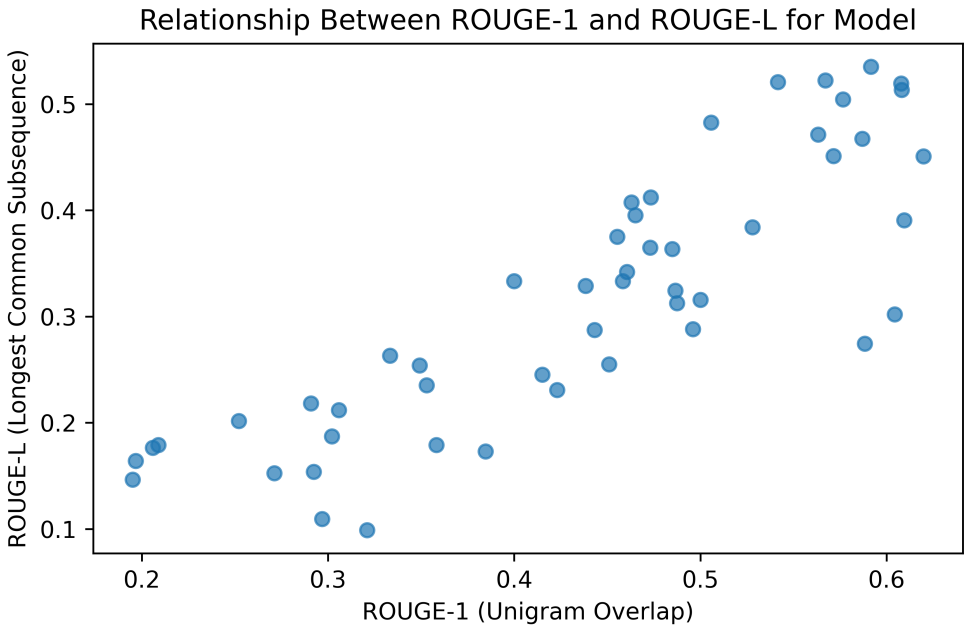

Figure 2: Relationship between ROUGE-1 and ROUGE-L scores for the model on the test set. Each point represents a single example.The shows a clear positive correlation between ROUGE-1 and ROUGE-L. When the model's summary shares more key words with the reference (higher ROUGE-1), it also tends to follow the reference's sentence structure more closely (higher ROUGE-L). This suggests that the model does not rely solely on keyword overlap, but also captures some aspects of summary organization.

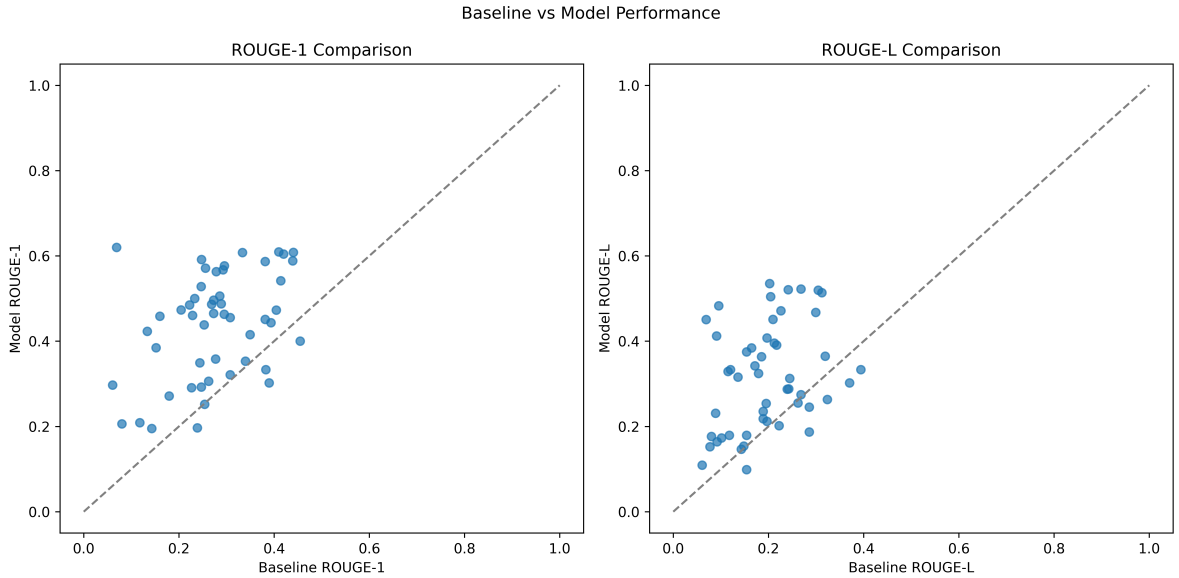

Figure 3: Comparison of baseline and model performance for ROUGE-1 (left) and ROUGE-L (right). Each point represents a single test example, and the diagonal line indicates equal performance between the two systems. In ROUGE-1, the model underperforms the baseline on 5 out of 50 examples. In ROUGE-L, the model underperforms on 9 examples, reflecting cases where the baseline's first sentence happens to align more closely with the human reference summary in structure or content.

Qualitative Examples

Below are examples comparing the baseline and model output.

Example 1

- **Document (truncated):** (CNN)Some of the men and women of the Indianapolis police force are giving up their blues. Beginning Friday, blue uniform shirts will be traded for white ones for command staff members of the Indianapolis Metropolitan Police Department (IMPD). In a statement, the department said the change is being made as part of its constant effort to ensure "accountability, professionalism and transparency... at the forefront of our day-to-day activities." As police departments around the country see more protests over the use of lethal force, IMPD officials acknowledge that this is a time of "increased scrutiny of police operations and tactics," but said the decision to change the uniform for certain ranks within the department is "not related to any specific, individual incident occurring elsewhere in ...
- **Gold summary:** Beginning Friday, some ranks of the Indianapolis police department will wear white shirts . Police say the change in attire is not related to any specific incident . The new uniform shirt color is aimed at ensuring accountability .
- **Baseline summary(R1F: 0.2264):** (CNN)Some of the men and women of the Indianapolis police force are giving up their blues.
- **AI pipeline summary(R1F: 0.2909):** The change is being made as part of an effort to ensure "accountability, professionalism and transparency," IMPD officials say .

Example 2

- **Document (truncated):** A Conservative party candidate has defended photographs which show him trying to lick a female friend's breasts saying they were taken on a night out with friends and show him being a normal, sociable person. Liam Marshall-Ascough is standing to be an MP in Stoke-on-Trent central, which had a 5,500 Labour majority in 2010, and is a traditional Labour stronghold. The candidate, who is gay, is pictured on the open profile of his Facebook page appearing to lick a female friend's breasts. Underneath the picture a friend commented 'drunko'. But Mr Marshall-Ascough has insisted the photos, taken seven years ago, just show him out having fun with close friends 'like a normal person' and do not undermine his ability to do the job. He also said they show he is able to interact with people - importa ...
- **Gold summary:** Liam Marshall-Ascough is a Tory candidate for Stoke-on-Trent Central . The Crawley councillor has been seen in pictures licking a friend's breasts . But he insists the photographs were taken seven years ago and show him having fun and being sociable, like a normal person in their twenties . Mr Marshall-Ascough said he enjoys interacting with people, which is important for a politician to not be 'just all about the politics'
- **Baseline summary(R1F: 0.4404):** A Conservative party candidate has defended photographs which show him trying to lick a female friend's breasts saying they were taken on a night out with friends and show him being a normal, sociable person.
- **AI pipeline summary(R1F: 0.0.6081):** Liam Marshall-Ascough is standing to be an MP in Stoke-on-Trent central . The 35-year-old is pictured on Facebook appearing to lick a female friend's breasts . Underneath the picture a friend commented 'drunko' He insists the photos, taken seven years ago, just show him having fun with friends 'like a normal person' He also said they show he is able to interact with people - important to the role of a politician .

Example 3

- **Document (truncated):** Manny Pacquiao has become one of the most recognisable stars in the world of sport after a series of mega-fights and another one to follow on Saturday night against Floyd Mayweather at the MGM Grand in Las Vegas. The Filipino has been involved in some of the biggest bouts in the world over the past decade or so, facing the likes of Oscar De La Hoya, Ricky Hatton, Miguel Cotto and four fights against Mexican warrior Juan Manuel Marquez. But it's been a tough road to the top for Pacquiao from being born into poverty in the Philippines. He spent the early part of his career boxing in the Far East before being given his chance on the big stage in 2001 against Lehlo Ledwaba in an IBF world super-bantamweight title fight in Las Vegas. Manny Pacquiao pictured as a teenager in a boxing gym in Mani ...
- **Gold summary:** Manny Pacquiao will take on Floyd Mayweather at the MGM Grand in Las Vegas on Saturday . The Filipino boxer is one of the biggest stars in the world of sport after a series of big fights . Pacquiao spent his early years boxing in Manila in the Philippines and was pictured training as a 17-year-old . Pictures taken in 1996 show Pacquiao training at the LM Gym in Manila before he became a star . The 36-year-old arrived in Las Vegas on Tuesday after a 270-mile journey from Los Angeles on his luxury bus . CLICK HERE for all the latest Manny Pacquiao vs Floyd Mayweather news .
- **Baseline summary(R1F: 0.3810):** Manny Pacquiao has become one of the most recognisable stars in the world of sport after a series of mega-fights and another one to follow on Saturday night against Floyd Mayweather at the MGM Grand in Las Vegas.
- **AI pipeline summary(R1F: 0.5870):** Manny Pacquiao will take on Floyd Mayweather at the MGM Grand Arena in Las Vegas on Saturday night . The 36-year-old is one of the biggest names in the world after a series of mega-fights . Pictures have emerged of the Filipino boxer as a teenager in a Manila boxing gym before he became a global superstar . The pair will make their first public appearances of the week on Tuesday as the bout draws closer .

5 Reflection and Limitations

This project demonstrated that even a lightweight summarization pipeline built with a pre-trained transformer model could outperform a simple heuristic baseline on average, achieving consistently higher ROUGE scores across the test set. The model worked better than expected in its ability to extract salient phrases and produce concise summaries without any fine-tuning, which confirms the strength of modern pretrained sequence-to-sequence models even in zero-shot conditions. However, several examples revealed unexpected failures: in some cases, the baseline's first sentence more accurately conveyed the key idea of the gold summary, yet the model still received a higher ROUGE score. For instance, in Example 1, the baseline correctly identifies the topic (Indianapolis police changing uniforms), whereas the model focuses on a secondary justification ("accountability, professionalism and transparency"), which is not the main point of the gold summary. ROUGE still assigned a higher score to the model output because it contained more overlapping high-frequency content words, showing that ROUGE does not always align with human judgments of summary quality. This highlights a core limitation of n-gram–based metrics: they evaluate lexical similarity rather than semantic alignment or factual correctness.[2] If given more time or computational resources, the next step would involve experimenting with embedding-based or LLM-based evaluation metrics (e.g., BERTScore or GPT-based evaluators)[3] , as well as fine-tuning a small summarization model to improve faithfulness and reduce cases where the model overgeneralizes or emphasizes irrelevant details.

References

[1] supkoon. Understanding ROUGE or Model Behavior (Korean Blog). https://supkoon.tistory.com/26, 2025.

[2] Adnan Masood. Recall-oriented understudy for gisting evaluation (rouge): The unseen metric that rules ai. https://medium.com/@adnanmasood/recall-oriented-understudy-for-gisting-evaluation-rouge-the-unseen-metric-that-rules-ai-5 2021. Accessed: 2025-XX-XX.

[3] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. URL https://arxiv.org/abs/1904.09675.