Muhammad Tanveer
Ansh Bhargava
CSCI 49362

NLP Project Baseline

Data

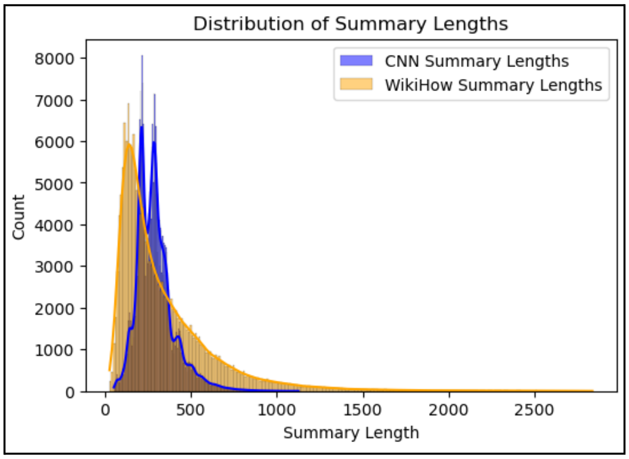
We use two datasets for this project, the CNN/DailyMail news article dataset and the WikiHow instructional guide dataset. These datasets are described further below.

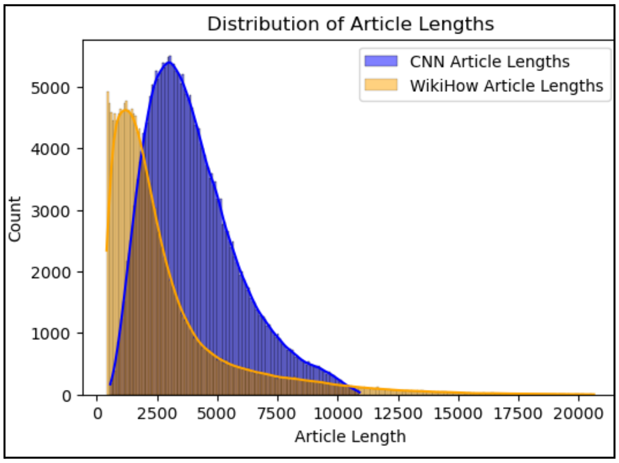
CNN/Dailymail (Source):
This is an English language dataset with over 300K unique news articles gathered from CNN and DailyMail publication sources. Each instance in the dataset contains the following features: ID (hex hash article source URL), article (string containing the body of the news article), highlights (string containing summary of article).

WikiHow (Source):
Large scale dataset sourced from WikiHow knowledge base. Each article is made up of multiple paragraphs with each paragraph starting with a sentence that summarizes the paragraph. The summary is constructed as a concatenation of the first sentence of each sub-section (each instruction).

We provide some basic summary statistics and distribution plots for these datasets:

Dataset Statistic	CNN/Dailymail	WikiHow
Number of article-summary pairs	311820	181925
Avg Length of Article	4018	2885
Avg Length of Summary	293	347
Summary to Article Ratio	0.088	0.204





For the baseline, we preprocessed our data using regex rules and other simple heuristics. This included replacing newlines with spaces, removing extra spaces before sentence ending punctuations, and removing particularly short articles. This was essentially to clean the data for better evaluation using the ROUGE metric.

For our final project, we will likely need to preprocess further, using tokenization and embeddings.

Baseline

Our baseline model for summarization was simply to select the first k sentences from the article as our summary. We also didn't add sentences if they were less than three words long. This was done in an attempt to minimize boilerplate sentences such as "By Associated Press" or "PUBLISHED" from being part of our summary. We implemented the baseline model for k between 2 and 6.

We chose this approach for its simplicity and because of the intuition that texts are structured to include salient information early on to hook readers. We anticipate the extent to which this is true to be domain specific.

Evaluation

Our evaluation metric was the ROUGE suite of scores. This is a popular metric for summarization evaluation and it measures the overlap of n-grams between the system and reference summaries, where n can vary (i.e. there exists ROUGE-1, ROUGE-2, etc). ROUGE-L measures the longest common subsequence. We obtained the following results where we bold the **best score** and underline the <u>second-best score</u> per metric.

CNN/DailyMail

Number of sentences used for baseline model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
2	27.77	10.74	17.75
3	32.65	13.17	20.80
4	**34.68**	14.45	<u>22.02</u>
5	<u>34.58</u>	**14.96**	**22.02**
6	33.48	<u>14.90</u>	21.51

WikiHow

Number of sentences used for baseline model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
2	23.54	5.32	15.41
3	25.46	5.95	**15.87**
4	**25.83**	6.25	15.71
5	25.60	6.41	15.37
6	25.13	**6.50**	15.00

We notice that around 4 is the number of sentences from the start of the article that seems to perform the best amongst our baseline models. However, the evidence is not particularly convincing for this.

We also notice that the scores for CNN/DailyMail are higher than WikiHow. This is intuitive since CNN/DailyMail are news articles which tend to put the most important information early in the article, whereas the WikiHow articles are instructional guides where the most important information is spread out throughout the article. This structural difference indicates that a better performing model would likely be able to discern importance beyond simple position in an article.

Baseline results from Notebook:
https://colab.research.google.com/drive/19LFmtLfEBPFLW8YyhXDmFNe9MuVkpb_y?authuser=2