NLP Project: Summarization

Muhammad Tanveer & Ansh Bhargava

Two fundamental approaches to summarization

- **Extractive** summarization is a method of summarization that involves **selecting and concatenating** the most important sentences or phrases from the original text to create a summary.

- **Abstractive** summarization is a method of summarization that involves **generating new sentences** that convey the meaning of the original text, rather than simply selecting and concatenating existing sentences.

Data

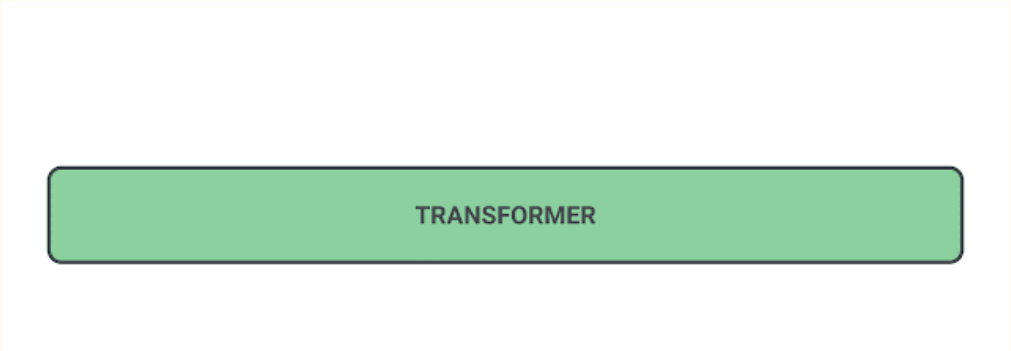
	Average article length	Average summary length	Average summary to article ratio
CNN/DailyMail dataset (news articles)	4018	293	8.8%
WikiHow dataset (how-to articles)	2885	347	20.4%

Baseline

Selecting the first k=5 sentences of the original text as the summary

PEGASUS overview (Abstractive)

- **P**re-training with **E**xtracted **G**ap-sentences for **A**bstractive **S**ummarization
- A self supervised model for abstractive text summarization
- Trained by masking of input text to produce the abstracted summary
- When tasked with abstractive summarization, it performs really well on major metrics.



https://ai.googleblog.com/2020/06/pegasus-state-of-art-model-for.html

BERTSUM overview (Extractive)

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- Bidirectionality uses words in both the previous and next context of the target word to create word embeddings
- Allows for masked inputs to be detected as well as relations between two concurrent sentences.
- BERTSUM builds on BERT by classifying whether or not sentences belong in the summary i.e. their level of importance in relation to the text

Extended summarizers

Graph Summarizer

Leverage TextRank algorithm to perform graph-based summarization.

1. Splits text into sentences and returns matrix of cosine similarities between each sentence
2. The textrank function then takes the list of sentences, builds the similarity matrix, and uses PageRank to score each sentence based on its importance.
3. Performs abstractive summarization on top (n=10) most ‘important’ sentences

Ensemble Summarizer

Leverages both extractive and abstractive summaries, ranking sentences in their outputs to generate final summary.

1. First extracts all sentences from both summaries and removing duplicates.
2. Similar to the graph summarizer, uses PageRank to score each sentence (across both summaries)
3. Returns the top (n=5)sentences with the highest PageRank scores

Other implemented summarization methods

- Simple two-step hybrid
  - First performs extractive summary on (long) source text. Then, return abstractive summary of first step's output.
- Length weighted two-step hybrid
  - Similar as above, but weights longer (proxy for importance) sentences in extractive summary more.
- Hierarchical
  - Breaks text into chunks (header / paragraph level) and summarizers each chunk. Concatenate each sub-summary.
- Iterative
  - Iteratively perform summarization, reducing length by merging 'redundant' sentences and removing short (proxy for uninformative) sentences.
- Query-based
  - More information retrieval: return a summary over 'relevant' sentences where relevancy is determined w.r.t user-given query

Metrics

The following metrics were used for to evaluate the text summarization:

- ROGUE-1, ROGUE-2, ROGUE-L
- Flesch Readability
- Entity Grid Score
- Jaccard Index

Results: CNN / DailyMail

	ROUGE-1	ROUGE-2	ROUGE-L	Readability	Entity Grid Score	Jaccard Index
Baseline	34.2	14.2	21.8	60.5	**0.488**	**0.193**
Abstractive	29.2	9.13	19.4	51.6	0.318	0.093
Extractive	**40.4**	**18.8**	**28.3**	**68.3**	0.382	0.105
Graph	28.9	8.00	18.6	51.9	0.352	0.090
Ensemble	36.0	15.9	22.0	64.5	0.477	0.149

Results: WikiHow

	ROUGE-1	ROUGE-2	ROUGE-L	Readability	Entity Grid Score	Jaccard Index
Baseline	24.3	**5.47**	14.7	70.0	0.208	**0.299**
Abstractive	21.1	3.99	14.2	58.3	0.178	0.086
Extractive	**25.1**	5.43	**15.8**	**76.1**	0.238	0.174
Graph	22.2	4.54	15.0	59.4	0.174	0.091
Ensemble	24.2	5.44	14.3	73.6	**0.255**	0.193

Questions?

Thank you!