

Identifying Factors Impacting News Summarization Performance of Pre-trained Transformer Models

...

Julia Ying & Ali A Nikooyan

NEWS Summarization

- Demand for automatically generated high quality news summaries on large-scale
- **Goal:** Identify the article-specific factors with the most impact on the quality of model generated summaries



Research Focus

- Importance of curriculum
- Importance of input length
- Importance of writing style

Dataset

CORNELL NEWSROOM

- 1.3 million articles and summaries
- 38 major publications
- variety of summarization strategies

Curriculum Quality

Max Grusky et al. 2018:

“We present NEWSROOM, a summarization dataset of 1.3 million articles and summaries written by authors and editors in newsrooms of 38 major news publications. Extracted from search and social media metadata between 1998 and 2017, these **high-quality summaries** demonstrate high diversity of summarization styles.”

Curriculum Quality

Max Grusky et al. 2018:

“We present NEWSROOM, a summarization dataset of 1.3 million articles and summaries written by authors and editors in newsrooms of 38 major news publications. Extracted from search and social media metadata between 1998 and 2017, these **high-quality summaries** demonstrate high diversity of summarization styles.”

Data Cleaning

- Foreign language

Kwamitin kula da wasannin Olympics na duniya, IOC, zai yanke kodai za a hana Russia shiga gasar wasannin Olympics ta Rio ko a a.

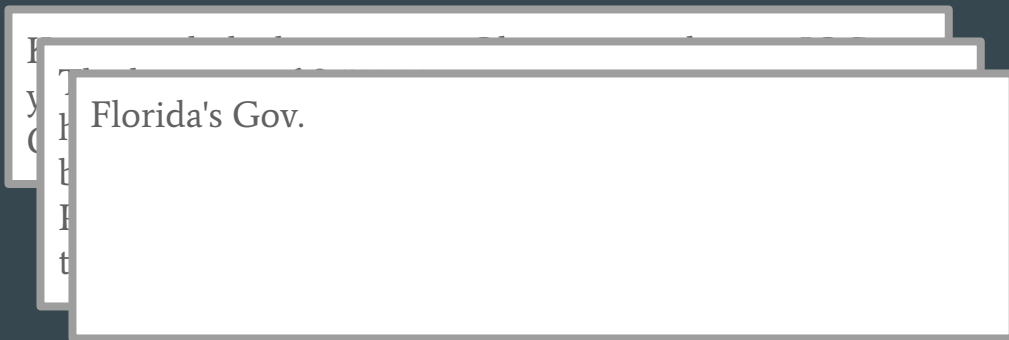
Data Cleaning

- Foreign language
- Urls in summaries

The banning of [Fuelling Poverty](http://www.youtube.com/watch?feature=player_embedded&v=udUBKu4go7s); leaves many Nigerians questioning the government of Goodluck Jonathan.

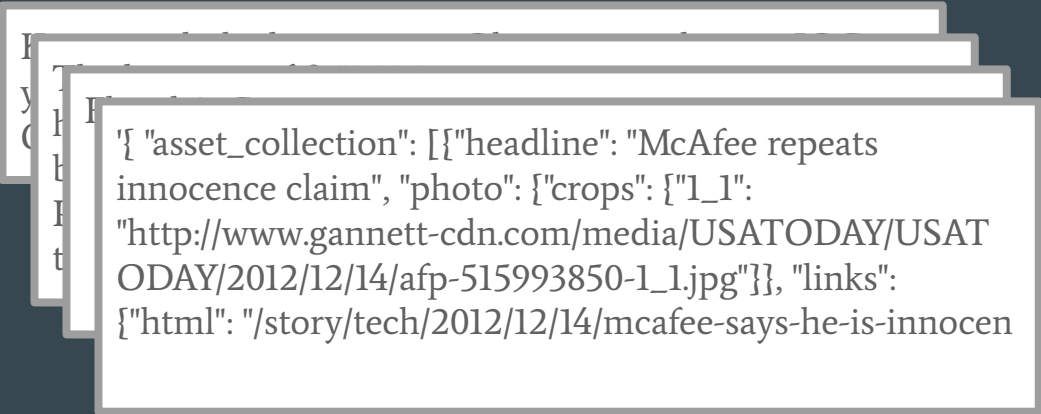
Data Cleaning

- Foreign language
- Urls in summaries
- Summaries too short



Data Cleaning

- Foreign language
- Urls in summaries
- Summaries too short
- Badly parsed



```
{ "asset_collection": [{ "headline": "McAfee repeats  
innocence claim", "photo": { "crops": { "1_1":  
"http://www.gannett-cdn.com/media/USATODAY/USAT  
ODAY/2012/12/14/afp-515993850-1_1.jpg" } }, "links":  
{ "html": "/story/tech/2012/12/14/mcafee-says-he-is-innocen
```

Data Cleaning

- Foreign language
- Urls in summaries
- Summaries too short
- Badly parsed
- Repeating N-grams

Oscars 2015: Red carpet 50 photos

Lady Gaga, left, and Keira Knightley

Oscars 2015: Red carpet 50 photos

Oscars 2015: Red carpet 50 photos

Oscars 2015: Red carpet 50 photos

Oscars 2015: Red carpet 50 photos

Oscars 2015: Red carpet 50 photos

Data Cleaning

- Foreign language
- Urls in summaries
- Summaries too short
- Badly parsed
- Repeating N-grams
- Other non-sense

Oscars 2015: Red carpet 50 photos

Lady Gaga, left, and Keira Knightley

Oscars 2015: Red carpet 50 photos

Oscars 2015: Red carpet 50 photos

Oscars 2015: Red carpet 50 photos

Oscars 2015: Red carpet 50 photos

Oscars 2015: Red carpet 50 photos

[illegible]

ARF ARF ARF ARF ARF GRRRR. ARF ARF.

[illegible]

It always seems like only the owners of the dog barking do not know how annoying the problem is with their answer to your complaint, "We never hear him bark" or, "He never bothers us"! 11:58 pm, the dog is barking. 1:07 am, yes, the dog is still barking. 2:42 am, still barking. 3:36 am, the dog is still barking. 4:44 am, that dog is still barking. 5:21 am, the dog still hasn't lost its voice. 6:02 am, damn that dog, it is still barking. 7:45 am, I have to get up in 15 minutes and haven't slept all night because that dog just won't stop barking. 7:58 am, the dog owners get up, silence at last (for at least an hour). 8:00 am beep beep beep beep beep beep, damn the alarm!!! Oh, well, I'll be back tonight for more of the same. :('

Biggest Problem

Exploitation of “inverted-pyramid” writing

If you're cutting back on stocks because interest rates are rising, you're making a mistake. But don't just take my word for it (after all, I am a dyed-in-the-wool dividend-stock fan). Ask Ned Davis Research, which released its latest research on the relationship between stocks and rates about a year ago. The [...]'

Biggest Problem

- Exploitation of “inverted-pyramid” writing
- Solution: ROUGE-2 recall of the reference summaries (RefSum ROUGE)
 - Reference: first 150 words of the article
 - Removed RefSum ROUGE > 0.15

Method

- ~600k records after data cleaning
 - 400k train, 100k val, 100k test
- Models:
 - PEGASUS
 - Finetuned on entire NEWSROOM
 - BART
 - Finetuned on CNN/DM
 - T5
 - Pretrained, no finetuning

Effects of Curriculum: Method

- ~600k records after data cleaning
 - 400k train, 100k dev, 100k test
- Models:
 - PEGASUS
 - Finetuned on entire NEWSROOM
 - BART
 - Finetuned on CNN/DM
 - T5
 - Pretrained, no finetuning

Effects of Curriculum: Method

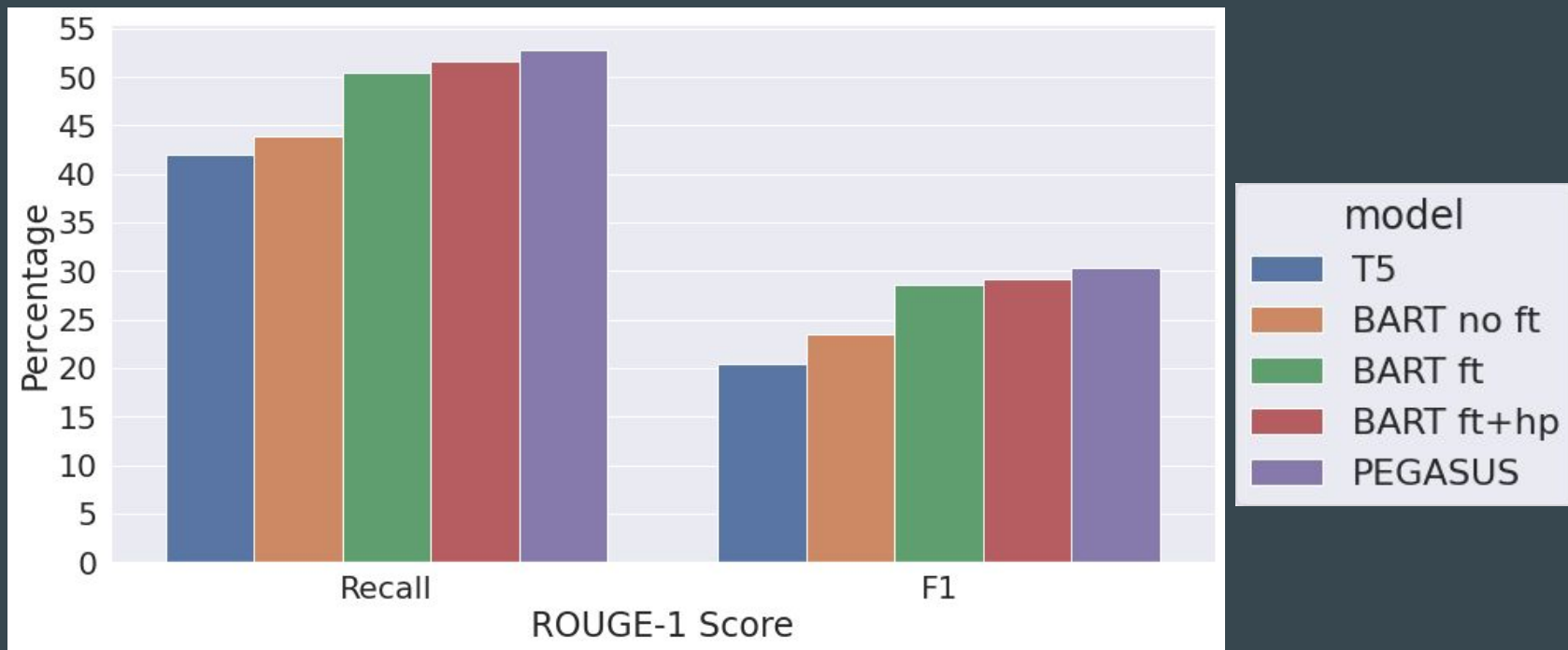
- ~600k records after data cleaning
 - 400k train, 100k dev, 100k test
- Models:
 - PEGASUS
 - Finetuned on entire NEWSROOM
 - BART
 - Finetuned on CNN/DM
 - T5
 - Pretrained, no finetuning

Test Performance

- T5
- PEGASUS
- BART no finetuning
- BART finetuned on cleaned NEWSROOM
- BART finetune on optimal hyperparameters

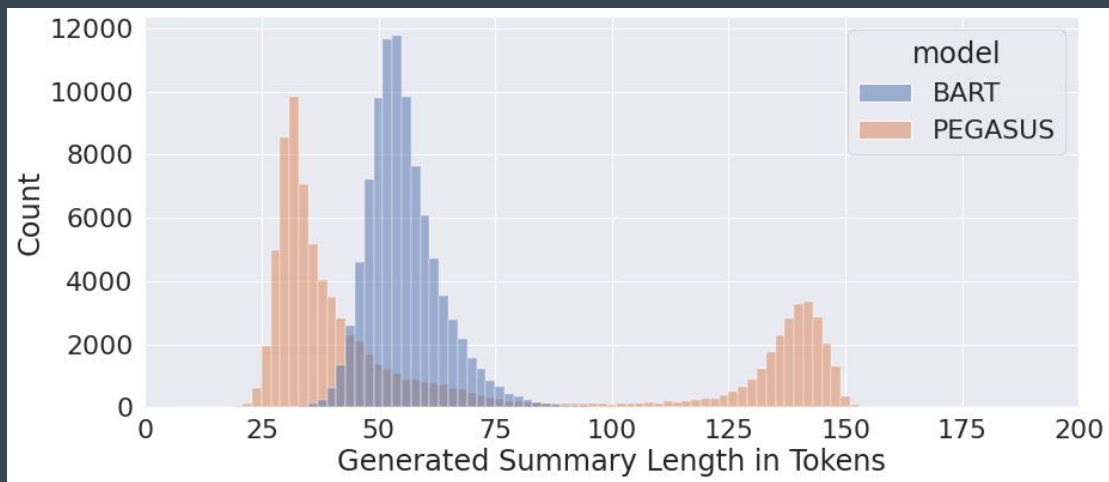
Q1: Effects of Curriculum

Q1

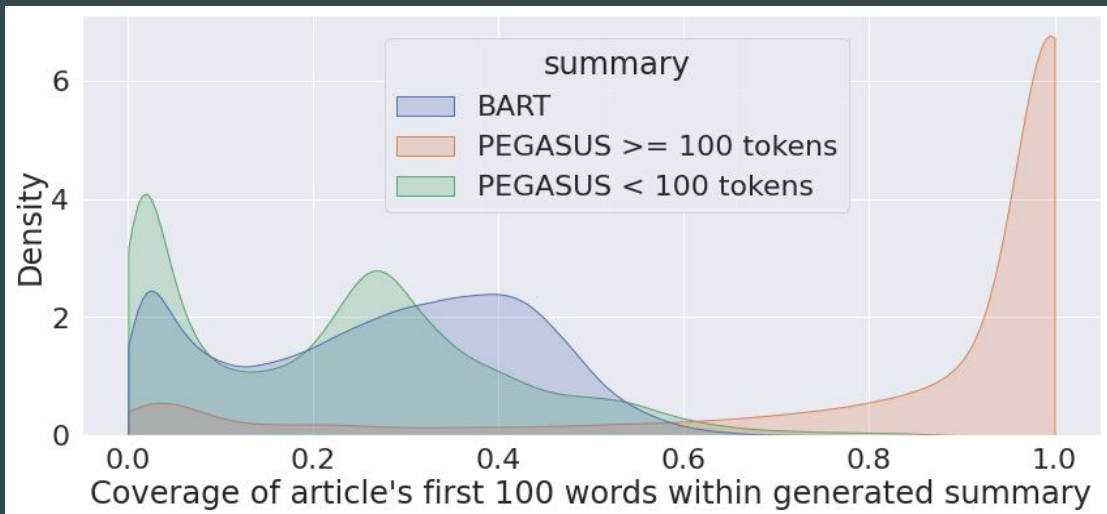
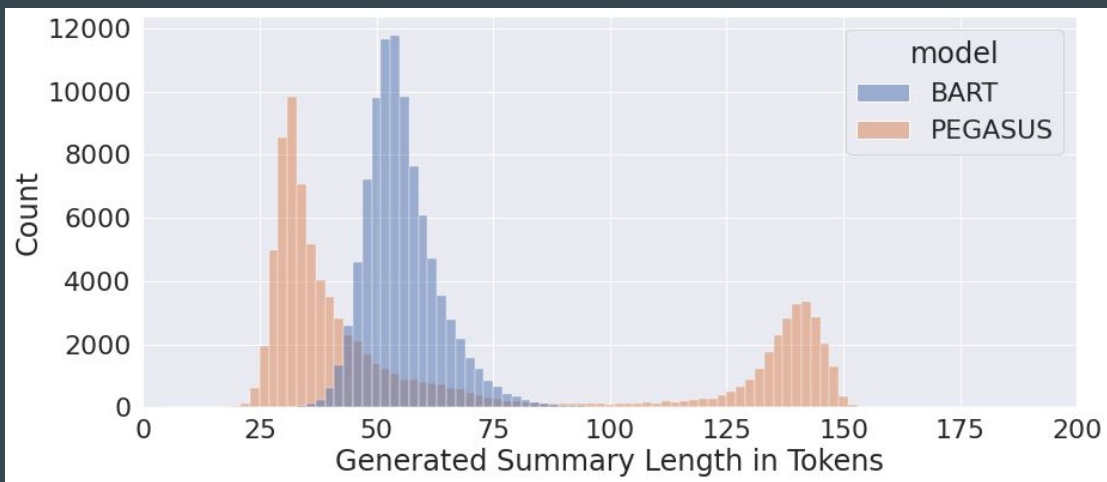


	Recall (R1/R2/RL)	F1 (R1/R2/RL)
PEGASUS	52.79/28.45/48.81	30.28/16.52/30.88
BART ft with optimal HP	51.67/26.72/46.81	29.13/14.50/29.14

Q1



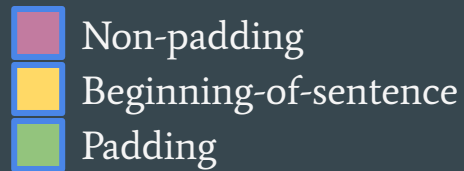
Q1



Q2: Effects of Truncation

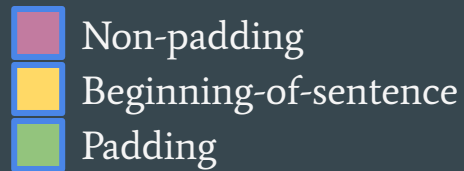
Q2 Method

Forward truncation



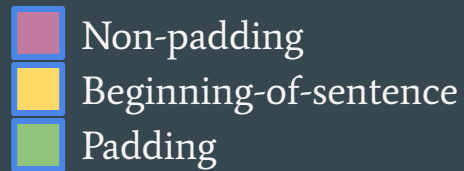
Q2 Method

Forward truncation



Q2 Method

Forward truncation



Q2 Method

Forward truncation



No trunc.



50% F trunc.



75% F trunc.



Non-padding



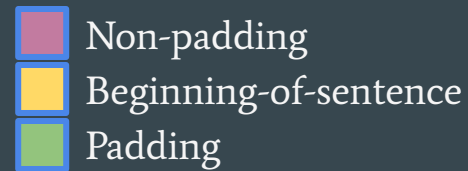
Beginning-of-sentence



Padding

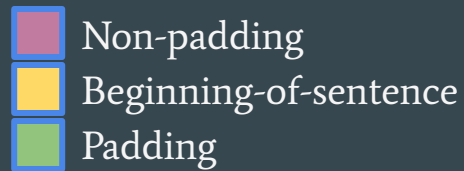
Q2 Method

Backward truncation



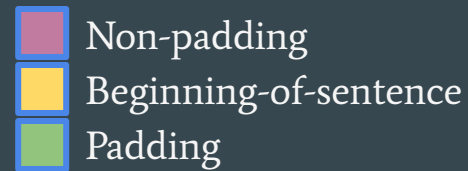
Q2 Method

Backward truncation



Q2 Method

Backward truncation



Q2 Method

Backward truncation



No trunc.



25% B trunc.



50% B trunc.



Non-padding

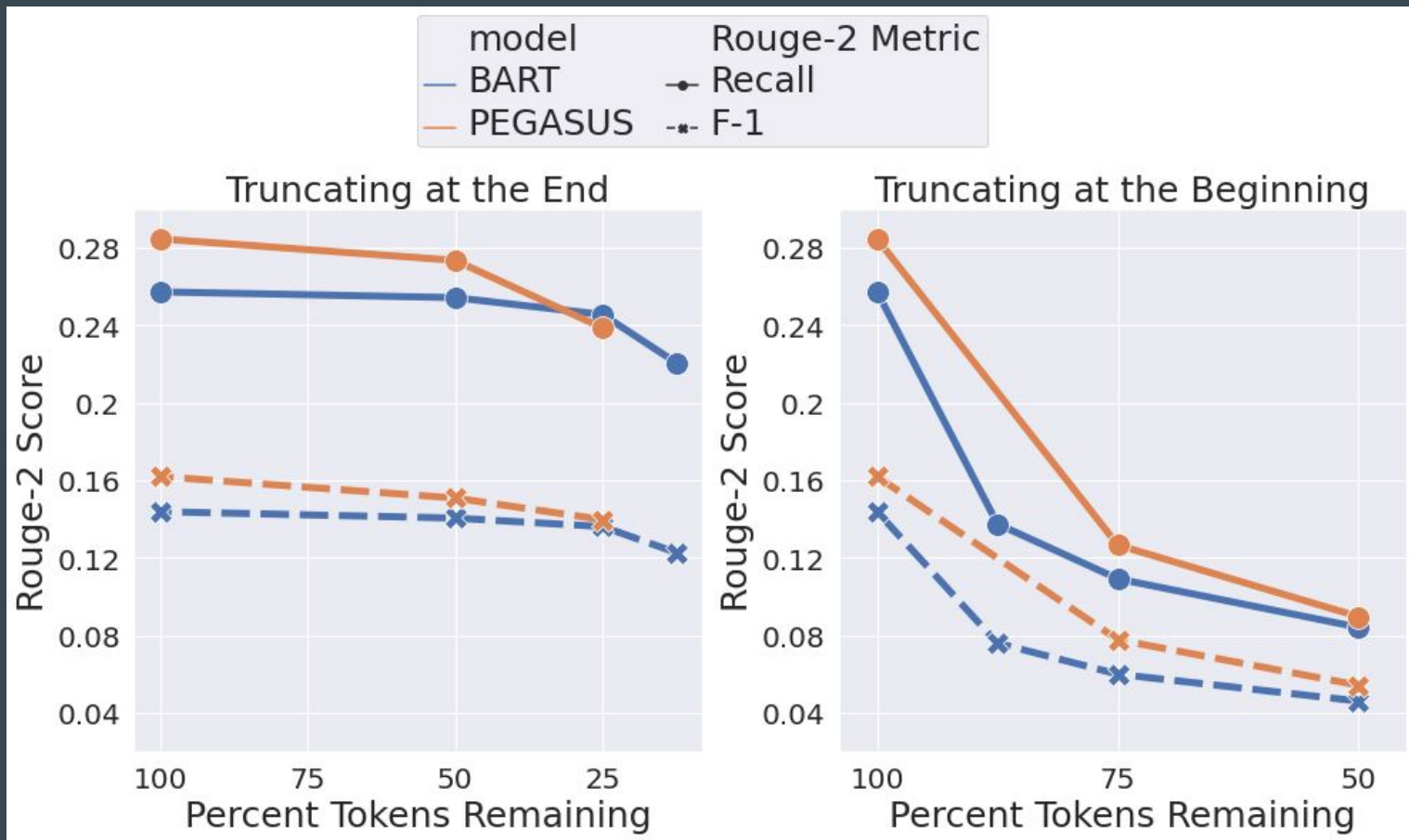


Beginning-of-sentence



Padding

Q2

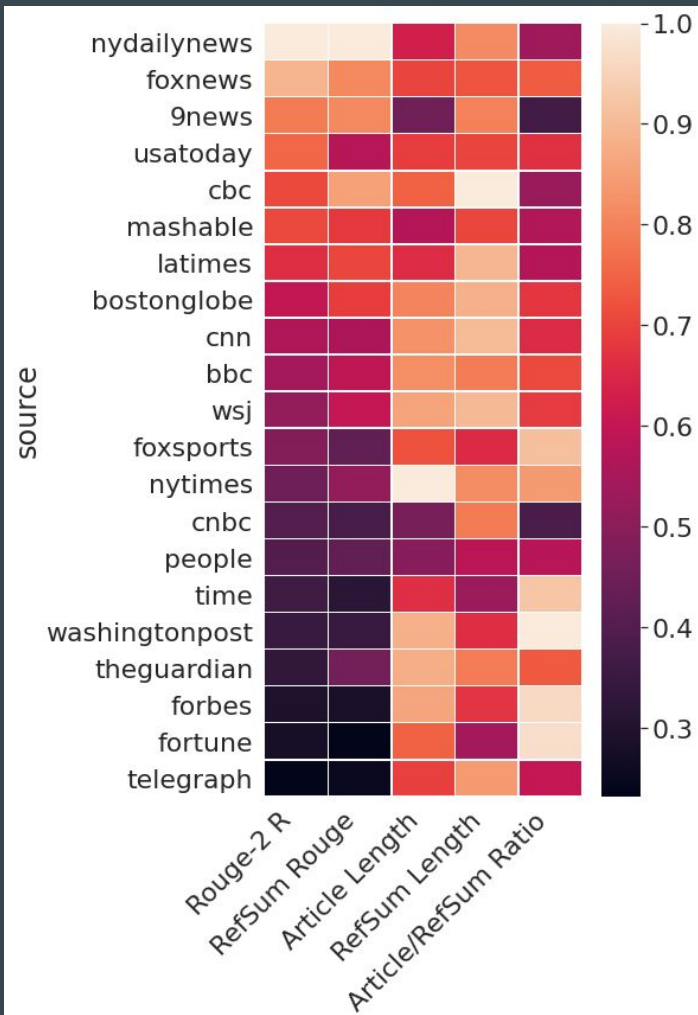


Q3: Effects of Writing Style

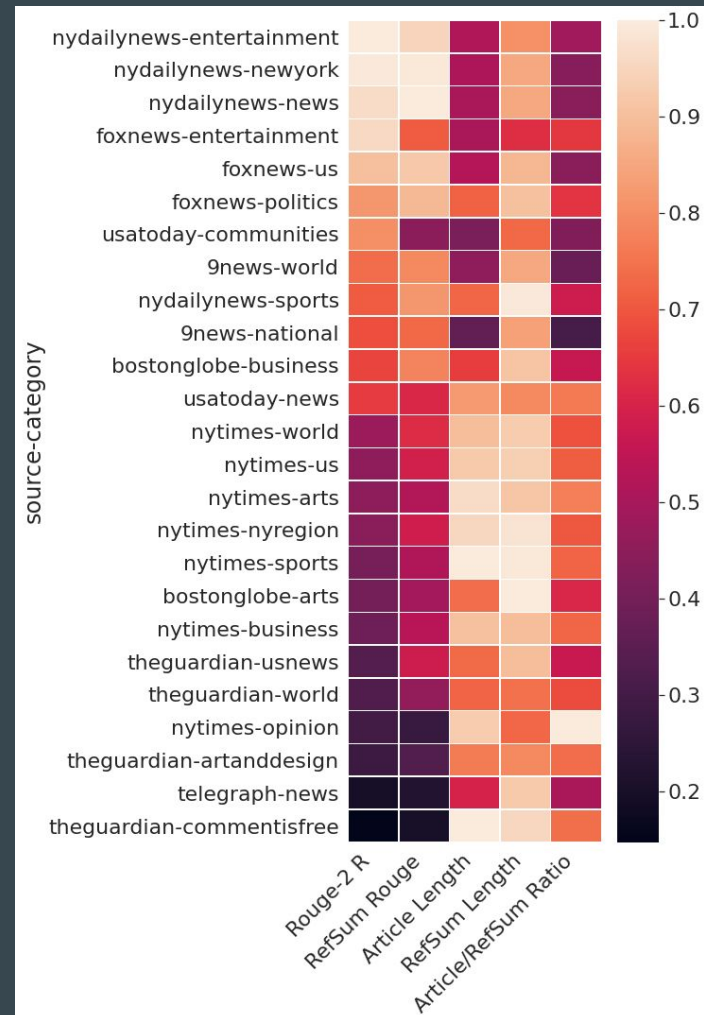
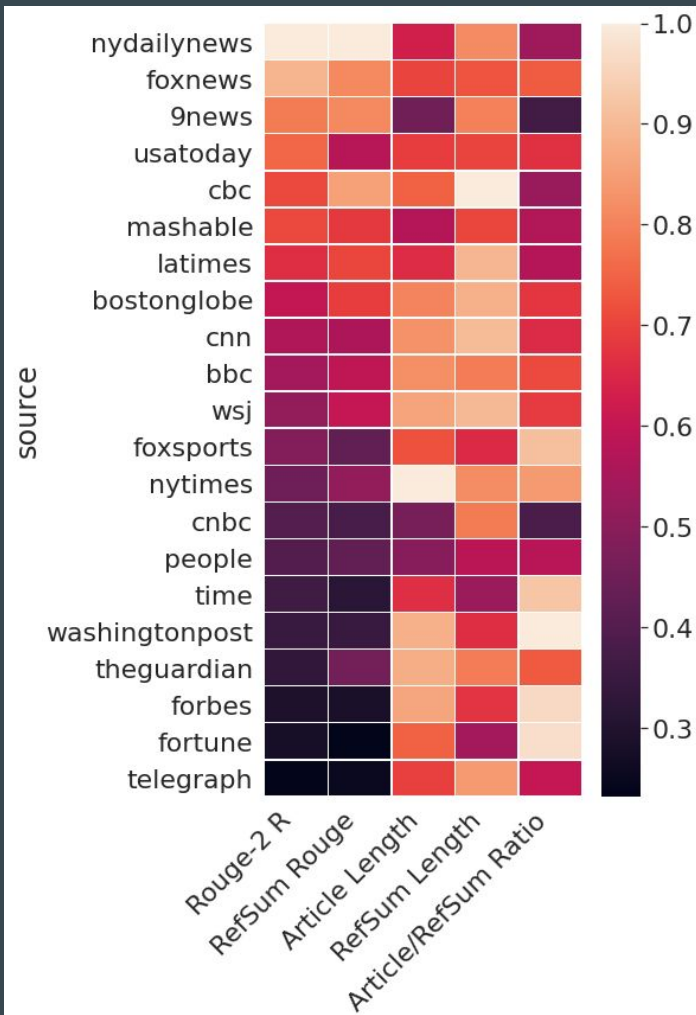
Potential factors to performance

- Article length
- Reference summary length
- Article/reference summary length ratio
- “Head-heaviness”
 - RefSum ROUGE

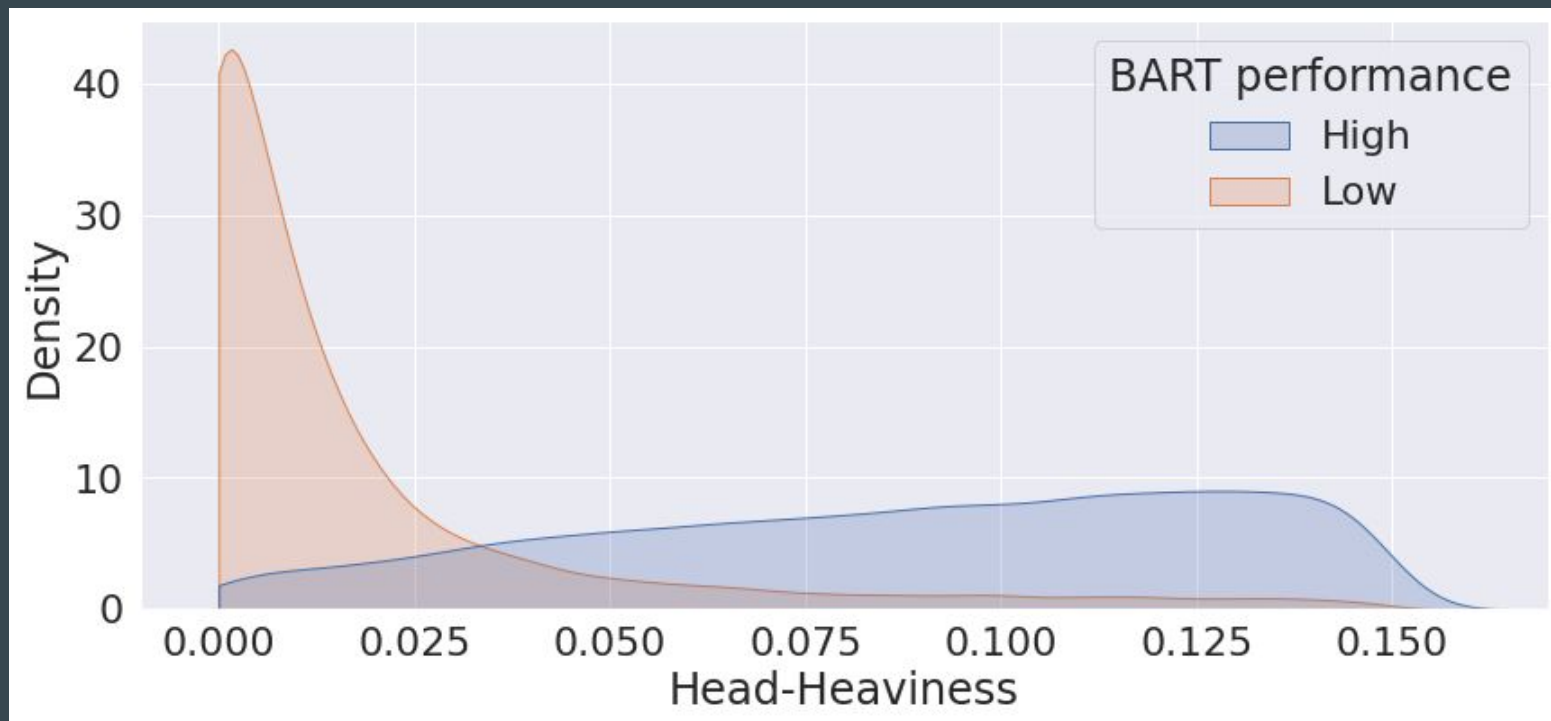
Q3



Q3



Q3



Conclusions

- Having a good curriculum is important
- Forward truncation by up to 50% had almost no effect on performance
- Inverted-pyramid writing style is critical to achieve good news summarization with BART

Conclusions

- Challenges/limitations:
 - ROUGE score: not the most accurate metric for “quality”
 - Limited model performance on articles not written in Inverted Pyramid style
- Future works:
 - Specialized abstractive summarization of news articles