CS5984: Big Data Text Summarization

Team 10: Electronic Theses and Dissertations

Ashish Baghudana, Beichen Liu, Guangchen Li, Stephen Lasky

December 4, 2018

Virginia Polytechnic Institute and State University
Blacksburg, Virginia - 24061
Instructor: Dr. Edward A. Fox

Dataset

ArXiv

Problem

Vocabulary and content of ETDs differs from news articles such as CNN and DailyMail.

Solution

Train summarization models with custom dataset containing academic content. We create a dataset from ArXiv:

- A repository of electronic preprints (known as e-prints) in the fields of mathematics, physics, astronomy, electrical engineering, computer science, etc.
- Assumption 1: Each conference article is approximately the same length as an ETD chapter
- Assumption 2: The abstract of a conference article approximates the summary of the article

Arxiv Dataset Characteristics

- Collected a subset of ~4,500 articles from ~41,000 entries
- Papers related to ML, CL, NER, AI, and CV field published between 1992 and 2018

Dataset	Avg. Text Length	Avg. Summary Length
CNN-DailyMail	~400 words	~100 words
ArXiv	6,009 words	178 words

- Significantly different compression ratios

ETD and ArXiv Preprocessing

One of the main challenges for creating the dataset was **parsing PDFs to text**. We tried two approaches:

GROBID

- GROBID (or Grobid) means GeneRation Of BIbliographic Data
- Machine learning library for extracting, parsing and re-structuring raw documents

Science Parse

- Science Parse parses scientific papers (in PDF form) and returns them in structured form (`JSON`)

Models

Model Summary

Trained Models

We trained three models – a baseline *seq2seq* model and two pointer generator (PGNs) models. The first PGN is trained with coverage disabled and the second PGN is trained with coverage enabled.

Pretrained Models

We also used a pretrained Fast Abstractive Summarization model (FastAbsRL) for generating summaries of the ArXiv dataset.

Model Evaluation

Evaluation

We divided the ArXiv dataset into train, test, and validation splits in a 70:15:15 ratio. We use the ROUGE metric to evaluate the models. We report these results on the test set of 707 articles.

Gold Standard Summary

Summary from a chapter of a master's thesis in Architecture

The last fifty years have seen a change in how and where Americans live and shop. Fewer people shop at Big Box stores such as Wal-Mart. As a result, several large sites and buildings have been left unoccupied. Communities have looked at bringing in new retailers and businesses into these buildings, as well as new types of development. However, new development often comes at the cost razing and reconstructing these sites. Instead of demolishing these buildings, they can be repurposed efficiently to avoid loss of investment.

Baseline seq2seq

Sample *seq2seq* Summary

last fifty years Americans have evolved in both how and where we live , and in how we shop for the things we need . Today we look at the things we need . Today we look at the things we need . Today we look at the things we need . Today we look at the things we need . Today we look at the things we need . Today we look at the things we need . Today we look at the

Baseline seq2seq

ROUGE-1, ROUGE-2, and ROUGE-L Results

Metric	Precision	Recall	F1-Score
ROUGE-1	0.2476	0.1178	0.1518
ROUGE-2	0.0214	0.0083	0.0112
ROUGE-L	0.2235	0.1041	0.1112

PGN without Coverage

Sample *PGN w/o* Coverage Summary

last fifty years Americans have evolved in both how and where we live , and in how we shop for the things we need . Today we look at the ubiquitous “ Big Box ” store from the past and see a way of shopping that fewer and fewer of us use regularly the While Realty division currently offers some 490 buildings and pieces of land for sale the

PGN without Coverage

ROUGE-1, ROUGE-2, and ROUGE-L Results

Metric	Precision	Recall	F1-Score
ROUGE-1	0.2325	0.2152	0.2150
ROUGE-2	0.0423	0.0380	0.0379
ROUGE-L	0.2087	0.1932	0.1827

PGN with Coverage

Sample PGN with Coverage Summary

Past few years and leaving communities wondering what to do with these large , imposing buildings the Wal-Mart Realty division currently offers some 490 buildings and pieces of land for sale the embodied energy these buildings already have invested into them from their construction . If we look for a new use without having to demolish.

PGN with Coverage

ROUGE-1, ROUGE-2, and ROUGE-L Results

Metric	Precision	Recall	F1-Score
ROUGE-1	0.2452	0.2216	0.2237
ROUGE-2	0.0431	0.0380	0.0382
ROUGE-L	0.2184	0.1979	0.1886

Fast Abstractive Summarization

Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting

- Abstractive / Extractive hybrid model with state-of-the-art results on the CNN / DailyMail datasets
- At a high level: extractively selects salient sentences and then abstractively rewrites (compresses) them
- ”Learns” how many sentences to extract (good for variable length)
- Tolerant of unforeseen terms

Challenges

- Has a very complex pipeline for training (on custom data) and testing. Appears to be fault intolerant. Created serious, time-consuming challenges with training on new data.

Sample Output on ArXiv

neonatal abstinence syndrome -lrb- nas -rrb- is commonly used to assist medical personnel . introduction to nas neonatal abstinence syndrome is a condition that can cause the newborn to go through withdrawal . the tool most widely used , as a basis for a scoring method . the project is aimed to be a step towards making a positive impact on the issue . for those that pharmacological treatment is deemed appropriate for , dosages are influenced by scores received . based on procedure published for ohio children 's ' hospitals , a score of 12 at two scorings .

Conclusions

Conclusions

- **Repetition of Words:** Baseline *seq2seq* and PGNs without coverage suffer from repetition of words.
- **Inhibitive Training Time:** PGNs took 80 hours to train.
- **Compression Ratios:** Different datasets differ in how they compress information. CNN-DailyMail has a compression ratio of 4:1 as compared to ArXiv which is at 40:1.

Future Work

Future Work

- **Hyperparameter Tuning:** Summarization tasks often do not converge to a global minimum without the right hyperparameters.
- **Human Evaluation of Summaries:** The ROUGE evaluation metric is not always indicative of the quality of the summary. Human evaluation would help estimate the quality of the summary more accurately.