CS 4984/5984
Big Data Text Summarization
Taught By: Edward A. Fox

Virginia Tech
Blacksburg, VA 24061

Abstractive Text Summarization of the Parkland Shooting Collection

Ryan Kingery, Jiacheng Ye, Li Huang, Chao Xu, Sudha Yellapantula

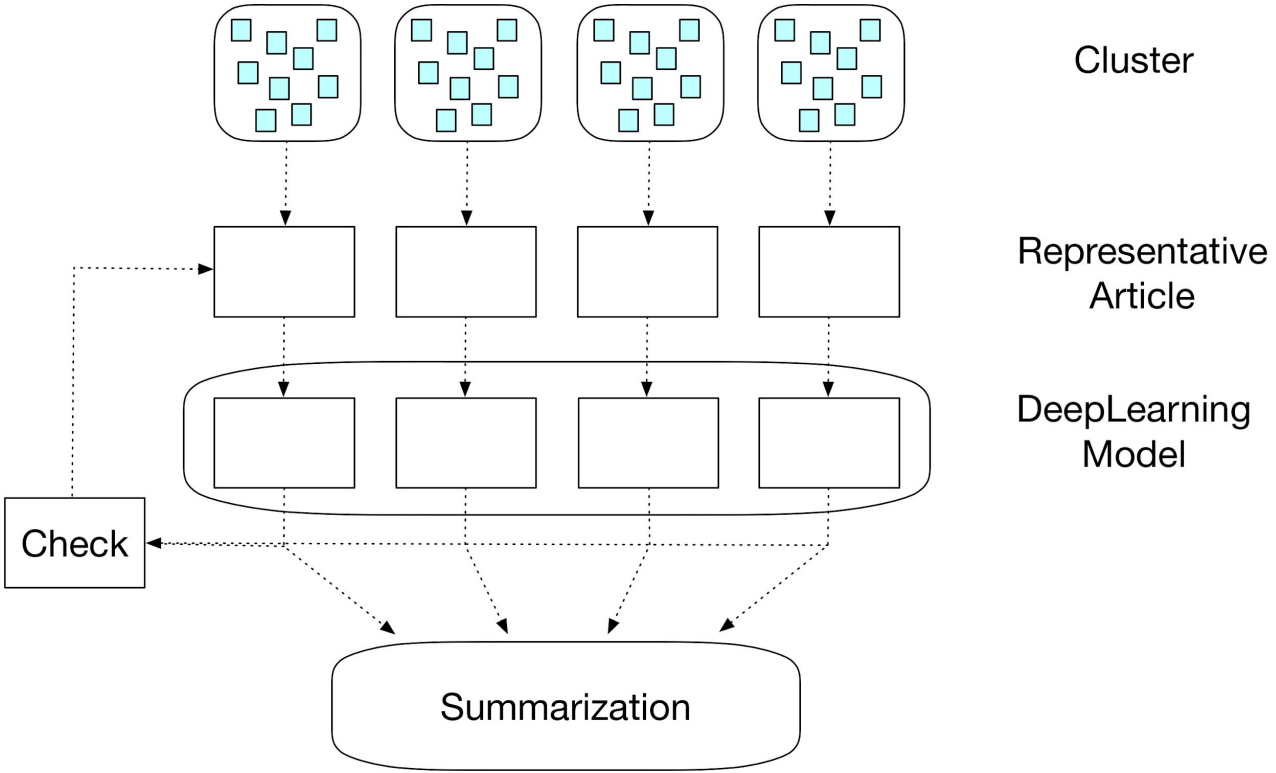
Submitted on December 10, 2018

Funding Source: NSF: IIS-1619028

Introduction

- Goal: Use deep learning techniques to generate a decent abstractive summary of the Parkland shooting web collection.
- Complications:
  - Dataset had to be converted from WARC file to text files.
  - Dataset was extremely noisy.
  - Denoised data was “too homogeneous” to easily segment.
  - Deep learning models only trained on individual articles, not collections.
  - Training necessary deep learning models from scratch was prohibitively slow.
  - Most of team had no real prior experience with ML, NLP, DL.

Collection Summarization Pipeline



Doc2Vec Embeddings + K-Means Clustering

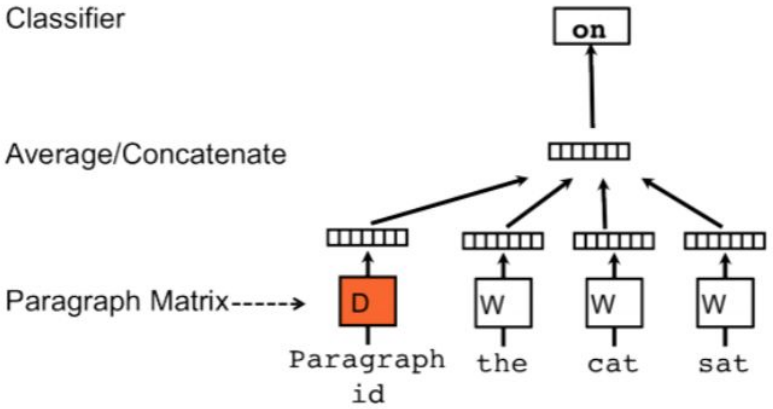
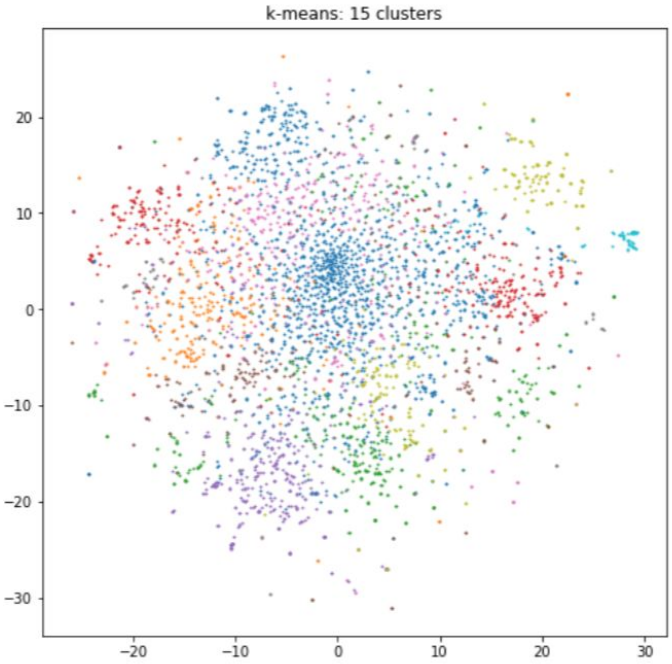


Figure 2.1: A framework for training Doc2Vec vectors. The document is combined with context words to predict a target word [1].



https://cs.stanford.edu/~quocle/paragraph_vector.pdf

Sequence-to-Sequence Model

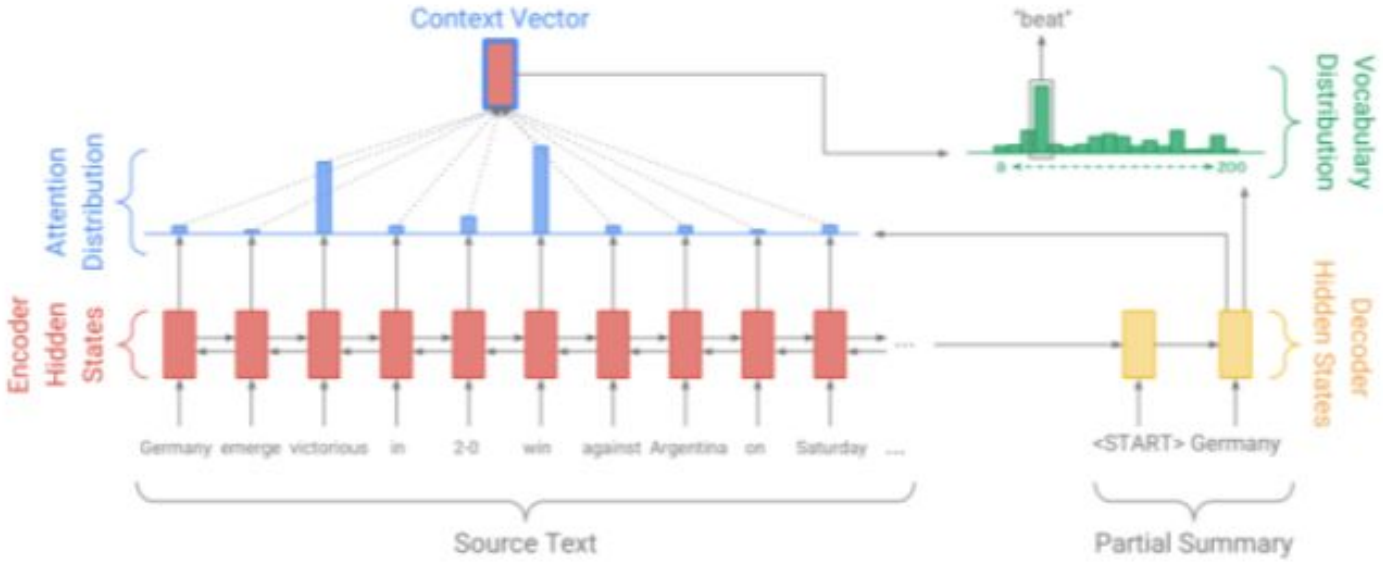


Figure 2.4: Illustration of a seq2seq model with attention.

https://arxiv.org/pdf/1602.06023.pdf

Pointer Generator Network

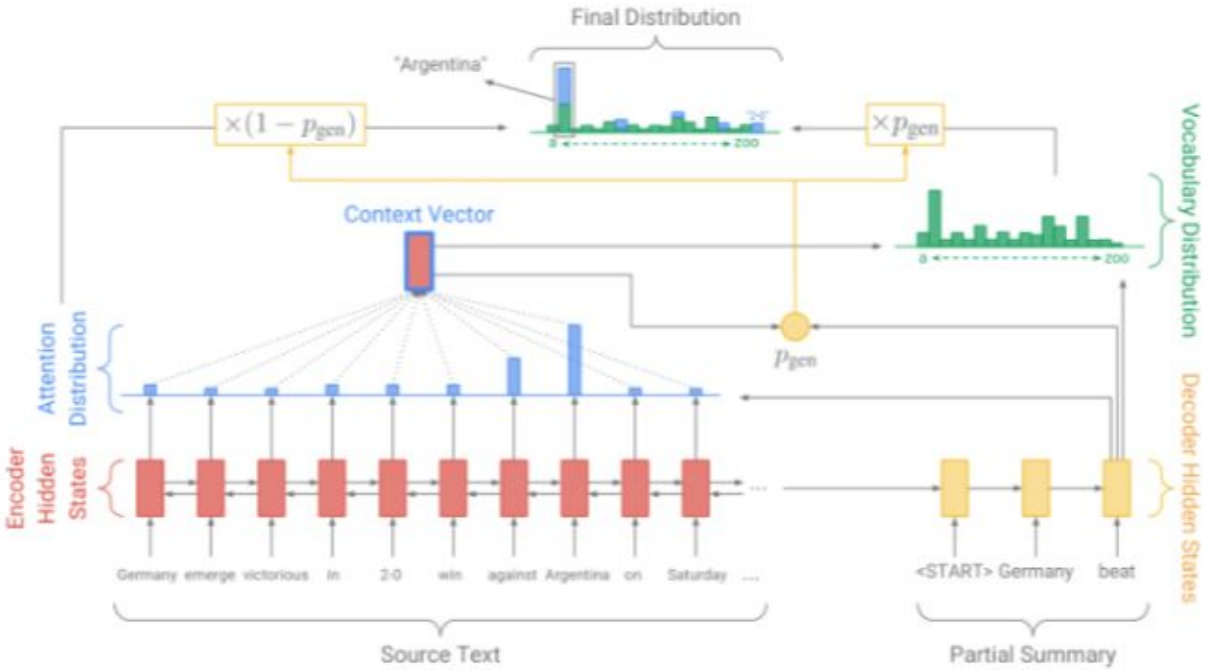


Figure 2.5: Illustration of a pointer-generator network.

https://arxiv.org/pdf/1704.04368.pdf

Reinforced Extractor-Abstractor Network

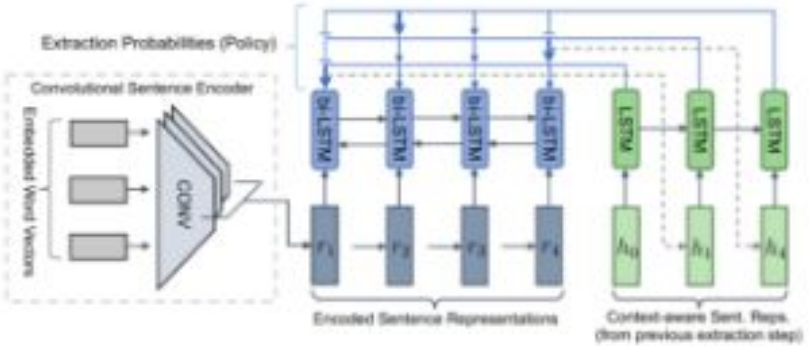


Figure 2.6: Illustration of the extractor agent [3]

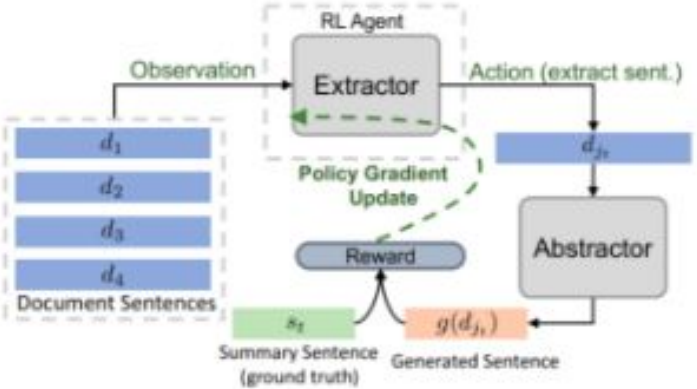


Figure 2.7: Reinforced training of the extractor (for one extraction step) and its interaction with the abstractor. [3]

https://arxiv.org/pdf/1805.11080.pdf

Example Representative Article

Presidential Proclamation Honoring the Victims of the Tragedy in Parkland, Florida: Our Nation grieves with those who have lost loved ones in the shooting at the Marjory Stoneman Douglas High School in Parkland, Florida. As a mark of solemn respect for the victims of the terrible act of violence perpetrated on February 14, 2018, by the authority vested in me as President of the United States by the Constitution and the laws of the United States of America, I hereby order that the flag of the United States shall be flown at half-staff at the White House and upon all public buildings and grounds, at all military posts and naval stations, and on all naval vessels of the Federal Government in the District of Columbia and throughout the United States and its Territories and possessions until sunset, February 19, 2018. I also direct that the flag shall be flown at half-staff for the same length of time at all United States embassies, legations, consular offices, and other facilities abroad, including all military facilities and naval vessels and stations. IN WITNESS WHEREOF, I have hereunto set my hand this fifteenth day of February, in the year of our Lord two thousand eighteen, and of the Independence of the United States of America the two hundred and forty-second.

Generated Summaries for Representative Article

The victims of the united states shall be flown at half-staff at the white house. The united states of america, i hereby order that the flag of the united states shall be flown at half-staff at the white house. The flag of the united states shall be flown at half-staff for the same length of time.

Seq2Seq

Presidential proclamation honoring the victims of the tragedy in parkland, florida. As a mark of solemn respect for the victims of the terrible act of violence perpetrated on february 14, 2018, by the authority vested in me as president of the united states of america, i hereby order that the flag of the united states shall be flown at half-staff at the white house and upon all public buildings and grounds, at all military posts and naval stations. In witness whereof, i have hereunto set my hand this fifteenth day of february, in the year of our lord two thousand eighteen years.

PGN

Use the transcript to help students with reading comprehension and vocabulary. The flag shall be flown at half-staff for the same length of time at all united states embassies, consular offices, and other facilities and naval vessels. He set my hand this fifteenth day of february, in the year of our lord two thousand eighteen. Presidential proclamation honoring the victims of the tragedy in parkland, florida our nation grieves with those who have lost loved ones.

REAN

Subjective Rank-Based Summary Evaluations

Table 3.2: Rankings for the generated summaries.

	Ryan	Sudha	Chao	Li	Jiacheng	Average
seq2seq-random	6	6	6	5	6	5.8
seq2seq-clustered	5	4	5	6	5	5.0
pgn-random	4	5	4	1	3	3.4
pgn-clustered	2	2	2	3	4	2.6
rean-random	3	3	3	2	2	2.6
rean-clustered	1	1	1	4	1	1.6

REAN with clusters was the best summary based on rankings alone.
Seq2Seq summaries were universally seen as bad.

Objective Summary Evaluations

Model	ROUGE-1	ROUGE-L	ROUGE-2	ROUGE-SU4
REAN-Clustered	0.09091	0.09091	0.03125	0.02198
REAN-Random	0.06061	0.0303	0.0	0.01099
PGN-Clustered	0.0303	0.0303	0.0	0.00549
PGN-Random	0.15152	0.09091	0.09375	0.04945
Seq2seq-Clustered	0.0	0.0	0.0	0.0
Seq2seq-Random	0.0	0.0	0.0	0.0

PGN with random articles was the best summary based on ROUGE-1, and gave the only ROUGE score that was higher than 0.1

Lessons Learned

- Abstractive summarization is still a long ways off before becoming anywhere near human-level summarization. Deep learning helps, but only superficially.
- Not all text datasets are easy to cluster semantically. Our dataset was nigh on impossible to do any coherent clustering or topic modeling on.
- Models trained on larger articles tend to give larger summeries. This suggests it would’ve perhaps been better to train on Wikipedia instead of CNN News.
- Using other people’s pretrained models was easier than training our own, but still wasn’t easy due to build and formatting issues.
- The REAN seems to capture more noise than the PGN, making PGN summaries seem more readable. Both are well above the Seq2Seq approach.