BDTS Final Presentation

Summarizing Hurricane Harvey

Team 1: Jack Geissinger, Theo Long, James Jung, Jordan Parent, Robert Rizzo
CS 4984/5984 Fall 2018 - 11/28/18

Address: Virginia Tech, Blacksburg, VA 24061
Instructor: Edward A. Fox

Presentation Outline

- **Preliminaries**
  - Data cleaning
  - Basic natural language processing
- **Text Summarization and Evaluation**
  - **Extractive Approaches**
    - TextRank
    - Templating
  - **Abstractive Approaches**
    - Pointer-generator network with Wikipedia
    - Multi-document pointer-generator networks with small dataset
    - Multi-document pointer-generator network with Wikipedia
- **Future work**

Cleaning our data for text processing

We discarded around 30% of our initial data with jusText.

The discarded information consisted of:

- HTML boilerplate content
  - Navigation links
  - HTML headers
  - Ads
- Unresponsive pages
- Non-English pages

Cleaning our data for text processing

We used a multi-threaded Python script that called on the jusText API and performed the following steps:

1. Set English stopword list from NLTK package
2. Call a URL from warc.gz file
3. Discard foreign language URLs and URLs that take too long to respond
4. Separate headings, paragraphs, and boilerplate content
5. Discard boilerplate content and sentences whose stopword densities are too high
6. Write to output JSON in following format
    a. {“URL”: url, “Title”:title, “Sentences”:sentences}
7. Repeat for all URLs

Basic natural language processing

Most frequent words

Tools: JSON Python Library and Natural Language Toolkit Library

- Step One: Concatenate all text from every article into one string
- Step Two: FreqDist Class
  - Outcome: Frequency of irrelevant stop words
- Step Three: NLTK Stopword corpus
  - Outcome: Stopwords filtered out
- Step Four: Python translator created using maketrans function
  - Outcome: Punctuation filtered out

Word	**Freq**
harvey	424
houston	422
texas	330
hurricane	311
said	214

Basic natural language processing

Part-of-speech tagging

Tools: JSON Python Library, Natural Language Toolkit Library, and TextBlob

- Step One: Make a TextBlob out of our sentences
- Step Two: Found POS tags from the TextBlob tags property
  - Outcome: All nouns and verbs identified with correct tense and plurality

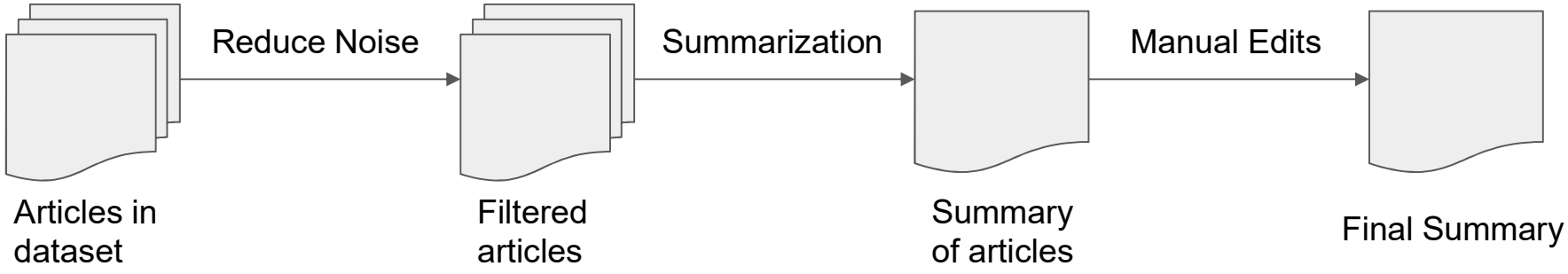
Noun	**POS Tag**
Houstonians	NNPS
Landsat	NNP
area	NN
storm	NN
Brazos	NNP

Verb	**Tag**
are	VBP
having	VBG
decide	VB
stays	VBZ
goes	VBZ

TextRank Summarization

We used the Summa library to generate a TextRank summary of our Hurricane Harvey data. [1]

We found that manual filtering and a limit of 100 words gives us the best result.



TextRank Summary Evaluation

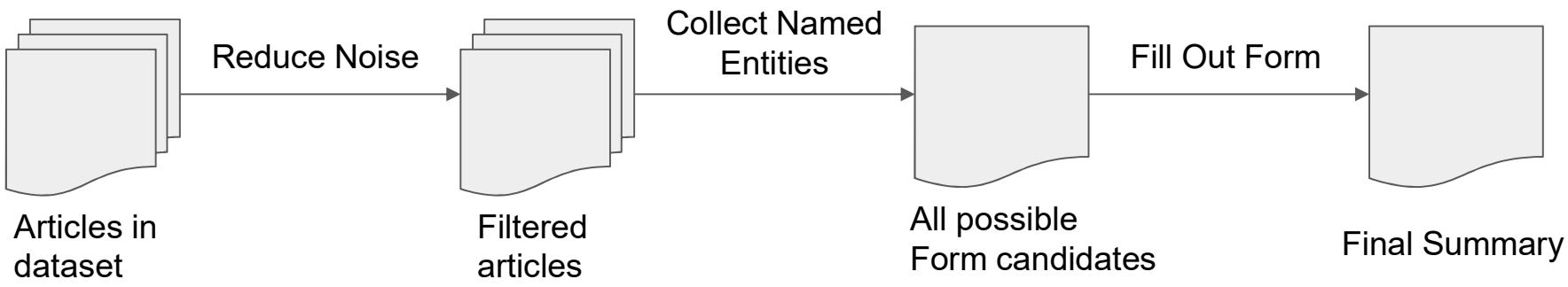
Our TextRank summary is human readable but lacks concrete figures and required manual editing.

"Houston has flooded each of the last three years, and while Harvey caused unprecedented damage in many neighborhoods, storms in 2015 and 2016 also displaced people and destroyed hundreds of homes. Teachers Volunteering in Shelters is a newly formed group of Houston-area teachers who are organizing to help children in the flood-ravaged areas of Southeast Texas brought on by Hurricane Harvey. Flooding unleashed by monster storm Harvey left Houston, the fourth-largest city in the United States, increasingly isolated as its airports and highways shut down and residents fled homes waist-deep in water."

ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
0.06897	0.0	0.06897	0.01266

Template Summarization

We used the spaCy library, and matplotlib library's hist function to generate a template summarization. [2]



Template Summary Evaluation

Our template summary turned out well with the slot values making sense:

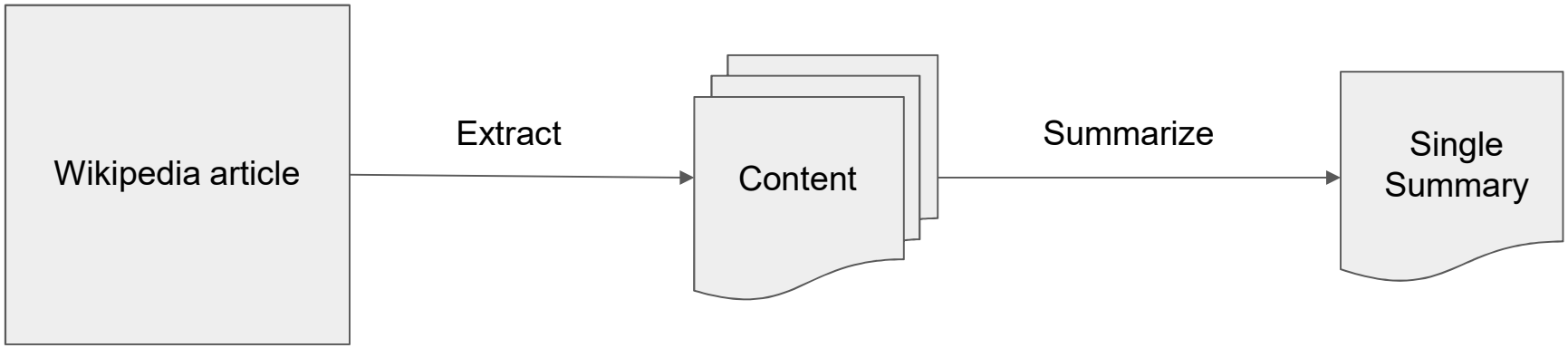
"Hurricane Harvey was a Category 4 hurricane that made landfall in August, 2017 in Houston. The hurricane traveled through the Gulf of Mexico with wind speeds that ranged from a low of 108 mph to a peak of 130 mph. In addition, the rainfall from Hurricane Harvey varied in areas, but there seemed to be around 45 inches to 53 inches in many areas. The organization FEMA was primarily involved with dealing with the affected area, and billions of dollars in damage was caused."

The ROUGE-1 and ROUGE-L scores are also good for an extractive summary:

ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
0.2069	0.07143	0.2069	0.05696

Pointer-generator Network with Wikipedia Article

We used the pointer-generated Network to summarize the Wikipedia article of Hurricane Harvey. [3], [4]



Pointer-generator Network Summary Evaluation

Our summary of the entire Wikipedia article using the pointer-generator network is readable, but slightly incoherent:

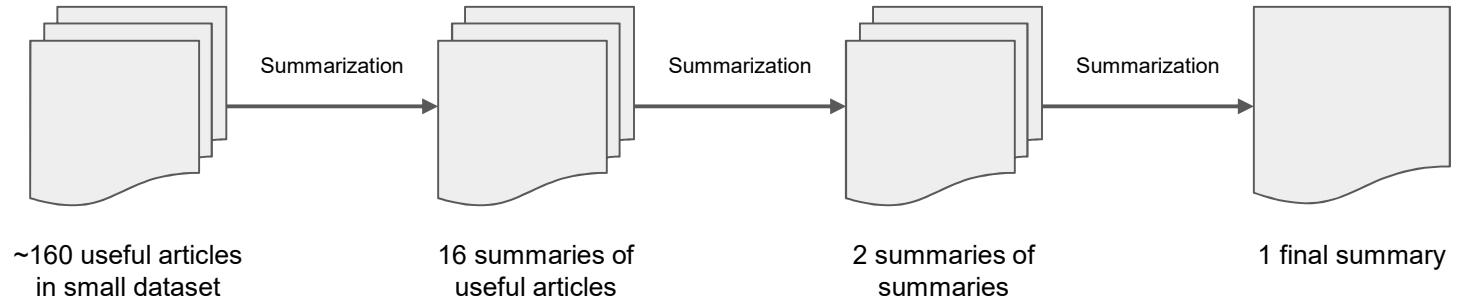
“Many areas received more than 40 inches of rain as the system slowly meandered over eastern Texas and adjacent waters. With peak accumulations of 60.58 in Nederland, Texas, Harvey was the wettest tropical cyclone on record, inflicting $125 billion in damage, primarily from catastrophic rainfall-triggered flooding in the Houston metropolitan area and Southeast Texas. It was the first major hurricane to make landfall in the United States since 2005, ending a record 12-year span in which no hurricanes made landfall at the intensity of a major hurricane throughout the country.”

Also, the ROUGE scores are very low and are the same as the TextRank approach:

ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
0.06897	0.0	0.06897	0.01266

Multi-document Summarization with Small Dataset

We found a GitHub repository that takes multiple articles and summarizes them with the pointer-generator network and the MMR algorithm. [5], [6]



This resulted in a coherent summary, but very low ROUGE scores.

Result from Multi-document Summarization of Small Dataset

The resulting summary is readable and coherent, which is quite incredible and unexpected:

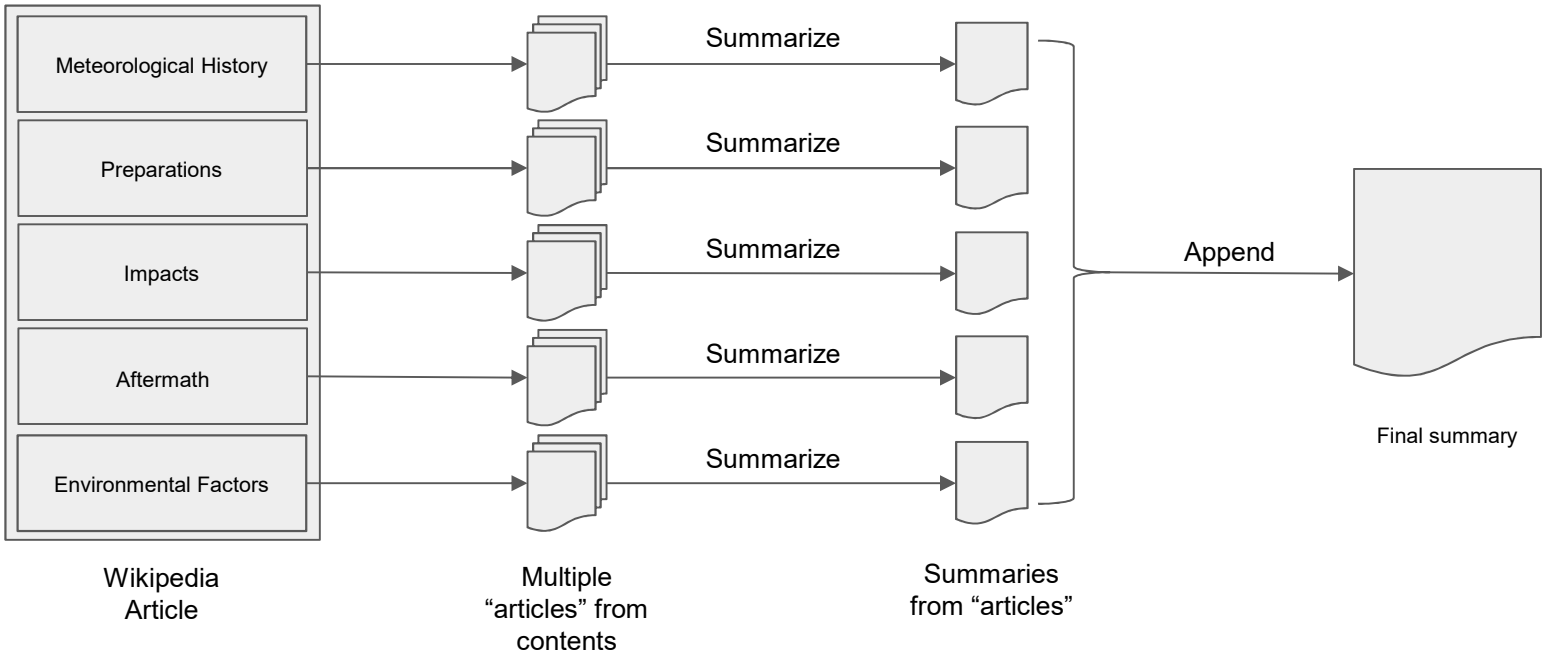
"Harvey is now the benchmark disaster of record in the United States. If past disasters are any indication, those numbers will only grow in the coming days and weeks. In fact, Harvey is likely already the worst rainstorm in U.S. history, but Harvey is in a class by itself. As a result, Harvey is the third 500-year flood to hit the houston area in the past three years, who have a federal flood policy in place. But a Category 4 hurricane and has lingered dropping heavy rain as a tropical storm."

But the ROUGE scores are very low:

ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
0.03448	0.0	0.03448	0.00633

Multi-document Summarization with Wikipedia Article

Another approach we tried, which worked very well, was summarizing the Wikipedia article for Hurricane Harvey using the GitHub library. [5], [6]



Result from Wikipedia Multi-document Summary

The summary is too long to post here, but the length is 487 words, which is a ~89% reduction from the original Wikipedia article.

In addition, the ROUGE scores look really good:

ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
0.51724	0.21429	0.34483	0.24684

This is our best summary by far.

Future Work and Conclusions

We think an improved pipeline for dataset summarization could help with analyzing large datasets, possibly with topic analysis and other methods.

We have shown that treating long documents as multiple documents can improve summarization accuracy compared to Gold Standards and could result in interesting research.

References

[1] R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Texts” [Online]. Available: University of Michigan, https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf. [Accessed November 26, 2018].

[2] D. Jurafsky and J. Martin, “Information Extraction,” in *Speech and Language Processing* [Online]. Available: Stanford, https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf. [Accessed September 27, 2018].

[3] A. See, “pointer-generator” [Online]. Available: GitHub, https://github.com/abisee/pointer-generator. [Accessed August 30, 2018].

[4] A. See, P. Liu, C. Manning, ‘Get To The Point: Summarization with Pointer-Generator Networks’, in *Association for Computational Linguistics,* 2017.

[5] L. Lebanoff, “multidoc_summarization” [Online]. Available: GitHub, https://github.com/ucfnlp/multidoc_summarization. [Accessed November 21, 2018].

[6] L. Lebanoff, K. Song, F. Liu, ‘Adapting the Neural Encoder-Decoder Framework from Single to Multi-Document Summarization’, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 2018.

Acknowledgments

Dr. Edward A. Fox, Email: fox@vt.edu

Liuqing Li, Email: liuqing@vt.edu

Chreston Miller, Email: chmille3@vt.edu

Global Event and Trend Archive Research (GETAR) funded by the National Science Foundation (IIS-1619028 and 1619371)