NEWSROOM

A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies

Max Grusky, Mor Naaman, Yoav Artzi

Cornell University

CORNELL
TECH
NEWSROOM OVERVIEW

A large, diverse summarization dataset.

- 1.3 million articles and summaries extracted from HTML metadata.
- Written in the newsrooms of 38 publications over the last 20 years.
- Wide range of both abstractive and extractive summaries.

extractive = borrows phrases from article
abstractive = uses more novel vocabulary

ABSTRACTIVE
South African photographer Anton Hammerl, missing in Libya since April 4th, was killed in Libya more than a month ago.

MIXED
A major climate protest in New York on Sunday could mark a seminal shift in the politics of global warming, just ahead of the U.N. Climate Summit.

EXTRACTIVE
A person familiar with the search tells The Associated Press that Texas has offered its head coaching job to Louisvilles Charlie Strong and he is expected to accept.
GOAL

Develop a large, diverse, and representative dataset to better understand the summarization task and to improve training of summarization systems.

OUTLINE

1. Comparing existing datasets
2. Collecting and developing NEWSROOM
3. Characterizing summary diversity
4. Developing a benchmark dataset
GOAL

Develop a large, diverse, and representative dataset to better understand the summarization task and to improve training of summarization systems.

OUTLINE

1. Comparing existing datasets
2. Collecting and developing NEWSROOM
3. Characterizing summary diversity
4. Developing a benchmark dataset
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Authors</th>
<th>Size</th>
<th>References</th>
<th>Sources</th>
<th>Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUC Datasets</td>
<td>Summary</td>
<td>NIST for DUC</td>
<td>Hundreds</td>
<td>Multiple</td>
<td>3</td>
<td>Abstractive</td>
</tr>
<tr>
<td>Gigaword Corpus</td>
<td>Headline</td>
<td>Journalists</td>
<td>9.9 million</td>
<td>Single</td>
<td>7</td>
<td>Abstractive</td>
</tr>
<tr>
<td>New York Times</td>
<td>Archival</td>
<td>Library scientists</td>
<td>600 thousand</td>
<td>Single</td>
<td>1</td>
<td>Extractive</td>
</tr>
<tr>
<td>CNN / Daily Mail</td>
<td>List of facts</td>
<td>Journalists</td>
<td>300 thousand</td>
<td>Single</td>
<td>2</td>
<td>Extractive</td>
</tr>
<tr>
<td>NEWSROOM</td>
<td>Summary</td>
<td>Journalists</td>
<td>1.3 million</td>
<td>Single</td>
<td>38</td>
<td>Both</td>
</tr>
</tbody>
</table>
GOAL

Develop a large, diverse, and representative dataset to better understand the summarization task and to improve training of summarization systems.

OUTLINE

1. Comparing existing datasets
2. Collecting and developing NEWSROOM
3. Characterizing summary diversity
4. Developing a benchmark dataset
SUMMARIZATION “IN THE WILD”

SNL / Hanks makes American dads great again as Baldwin’s Trump tics on
As a bizarre election drew near, a familiar face showed how to guest-host. Kate McKinnon’s inspired Clinton, meanwhile, got the knives out for The Donald

Sshhh! / How the cult of quiet can change your life
Silent retreats, silent restaurants and even silent dating events are on the rise. Now a new film aims to - quietly - spread the word

At Least 13 Dead After Bus and Truck Crash in California
At least 13 people are dead following a tour bus and semi-truck collision in Southern California. The bus crashed into the back of the truck on Interstate 10 north of Palm Springs, the California Highway Patrol said

Back-to-Back Explosions Hit Japanese City
Two apparent explosions hit the Japanese city of Utsunomiya, killing one person and injuring three others in what police are viewing as a possible suicide. The victim, believed to be a 72-year-old former military officer, may have set his house on fire

Cubs Advance to World Series
The Chicago Cubs took their first pennant since 1945, beating the Los Angeles Dodgers 5-0 Saturday night to win the National League Championship Series in six games. The team will face the Cleveland Indians in the World Series

AT&T’s Deal to Buy Time Warner Faces Political Barrage
AT&T’s blockbuster $85.4 billion deal to buy Time Warner promises to reshape the media landscape—if the companies can navigate a series of obstacles, including possible opposition from U.S. antitrust authorities and objections by lawmakers and media and telecom rivals. 8 minutes ago
• The Making of the Deal
• Heard: AT&T Dreams of a Hollywood Ending

The Fatal Mistake That Doomed Samsung’s Note 7
On the verge of challenging Apple’s global dominance in mobile phones, the South Korean company made a rushed decision, based on incomplete evidence, that would later force it to kill the model. 98

Iowa Governor Is Trump’s Pick for China Ambassador
By CHRIS BUCKLEY, MAGGIE HABERMAN and NICHOLAS FANDOS
24 minutes ago
An Iowa governor might seem an unlikely choice for one of the more sensitive diplomatic posts, but Gov. Terry Branstad has ties with China that go back decades.
• Romney Still in the Running for Secretary of State, Trump Says 42 minutes ago

Trump’s 5 Easiest and 5 Most Difficult Promises to Keep
By BINYAMIN APPELBAUM and ROBERT PEAR
5:11 AM ET
Mr. Trump made many pledges on the campaign trail. Here’s a look at where he’ll have the least, and the most, resistance.
The Fatal Mistake That Doomed Samsung’s Galaxy Note

By Jonathan Cheng in Seoul and
John D. McKinnon in Washington

Oct. 23, 2016 2:56 p.m. ET

The X-ray and CT scans showed a pronounced bulge.

After reports of Galaxy Note 7 smartphones catching fire spread in early September, Samsung Electronics Co. executives debated how to respond. Some were skeptical the incidents amounted to much, according to people familiar with the meetings, but others thought the company needed to act decisively.

High-quality social, search, and onsite summaries are encoded in article metadata.
DEVELOPING NEWSROOM

1. Preliminary scraping
2. Archive.org index/path scraping
   Text and metadata extraction
3. Automatic summary filtering
   (full details in paper)
4. Summary style analysis and partitioning
   (full details also in paper)
5. 1,000 top-ranked Alexa.com online publications
6. 38 publications using human summaries
7. 100+ million article-summary pairs
   (including automatic summaries and extraction errors)
8. 1.3 million articles and paired human-written summaries
9. 1.3 million annotated articles and summaries
   (labeled by date, publication, extractiveness, etc.)
GOAL

Develop a large, diverse, and representative dataset to better understand the summarization task and to improve training of summarization systems.

OUTLINE

1. Comparing existing datasets
2. Collecting and developing NEWSROOM
3. Characterizing summary diversity
4. Developing a benchmark dataset
PUBLICATION DIVERSITY

Major US Newspapers
PUBLICATION DIVERSITY

Financial Writing
We develop three helpful measures to understand human summarization style:

- **Compression**: How much smaller are summaries compared to their article?
- **Coverage**: What percentage of the summary words are from the article?
- **Density**: How long are the extractive fragments used by a summary?

Can we use these measures to better understand dataset diversity?

\[
\text{Compression}(A, S) = \frac{|A|}{|S|}
\]

\[
\text{Coverage}(A, S) = \frac{1}{|S|} \sum_{f \in \mathcal{F}(A, S)} |f|
\]

\[
\text{Density}(A, S) = \frac{1}{|S|} \sum_{f \in \mathcal{F}(A, S)} |f|^2
\]
South African photographer Anton Hammerl, missing in Libya since April 4th, was killed in Libya more than a month ago.

A person familiar with the search tells The Associated Press that Texas has offered its head coaching job to Louisvilles Charlie Strong and he is expected to accept.

Density is used as the primary measure of extractiveness.
**DIVERSITY OF EXTRACTIVE STYLES**

NEWSROOM extractive diversity consists of summarization style variation both within and between publications.

(see paper for detailed figures)
NEWSROOM has the widest range of extractive diversity of common summarization datasets.

(see paper for detailed figures)
GOAL
Develop a large, diverse, and representative dataset to better understand the summarization task and to improve training of summarization systems.

OUTLINE
1. Comparing existing datasets
2. Collecting and developing NEWSROOM
3. Characterizing summary diversity
4. Developing a benchmark dataset
We want to use NEWSROOM to evaluate different systems:

- Standard automatic evaluation with ROUGE.
- Multi-dimensional human evaluation.

But we also want to evaluate NEWSROOM:

- Is NEWSROOM useful for training systems?
- Does training on NEWSROOM improve system performance?
**EVALUATING SUMMARIES WITH ROUGE**

Is NEWSROOM useful for training systems?

**Pointer Model (See et al., 2017)**

*High-performing model combining extraction and abstraction.*

- Trained on CNN / Daily Mail
- Trained on NEWSROOM Subset
- Trained on NEWSROOM

\{ same size \}

- NEWSROOM systems outperform the CNN / Daily Mail system on out-of-domain DUC summaries.
- The NEWSROOM subset system outperforms the CNN / Daily Mail system on its own training data.
• ROUGE is not the complete picture.

• Human evaluation: important component of benchmark dataset evaluation protocol.

• Four-dimensional rating task capturing both semantic and syntactic aspects of summarization.

• Extension of evaluation concepts from Paulus et al. (2017) and Tan et al. (2017).
**HUMAN EVALUATION — SEMANTIC/SYNTACTIC**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pointer (NEWSROOM)</th>
<th>TextRank (unsupervised)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informativeness</td>
<td>3.77</td>
<td>3.61</td>
</tr>
<tr>
<td>Relevance</td>
<td>4.02</td>
<td>3.92</td>
</tr>
<tr>
<td>Fluency</td>
<td>3.87</td>
<td>3.92</td>
</tr>
<tr>
<td>Coherence</td>
<td>3.86</td>
<td>3.56</td>
</tr>
</tbody>
</table>

**Pointer (NEWSROOM)**

Combined abstractive and extractive, writes summaries token-by-token.

**TextRank (unsupervised)**

Fully extractive, write summaries by picking sentences from the article.

- Pointer better captures importance, relevance of ideas in a text.
- However, TextRank performs better in writing clarity and grammaticality.
• Full data and code are available online.

• Additional data analysis, system evaluation, and example summaries.

• Unreleased, private test set and leaderboard.

---

**THE DATA**

CORNELL NEWSROOM is a large dataset for training and evaluating summarization systems. It contains 1.3 million articles and summaries written by authors and editors in the newsrooms of 38 major publications. The summaries are obtained from search and social metadata between 1998 and 2017 and use a variety of summarization strategies combining extraction and abstraction.

---

**GETTING STARTED**

Use this site to explore the dataset and better understand the task of summarization as used by newsrooms around the Web.

Explore example summaries in the dataset across publications, time, and summarization strategies, analyze overall statistics in the dataset across these categories, and evaluate performance of existing summarization systems trained and tested on the unreleased NEWSROOM data.

The full dataset is available to download online with tools for extracting text and summaries from Archive.org, analyzing summary extractiveness, and evaluating system performance.

---

**SUBMISSION**

NEWSROOM contains an independent, unreleased test dataset for evaluation of leaderboard submission. To submit a system for evaluation, read the submission instructions. To preserve the integrity of the unreleased test data, submissions must be made at least two months apart.

---

**LEADERBOARD**

<table>
<thead>
<tr>
<th>Date</th>
<th>System</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
<th>HUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 Jan 2018*</td>
<td>Lede-3 Baseline (Gusky et al., 2018)</td>
<td>30.63</td>
<td>21.41</td>
<td>28.57</td>
<td>2.42</td>
</tr>
<tr>
<td>20 Jan 2018*</td>
<td>TextRank (Barries et al., 2016)</td>
<td>22.76</td>
<td>9.80</td>
<td>18.97</td>
<td>2.32</td>
</tr>
<tr>
<td>20 Jan 2018*</td>
<td>Seq2Seq + Attention (Bush et al., 2015)</td>
<td>5.91</td>
<td>0.43</td>
<td>5.36</td>
<td>1.93</td>
</tr>
</tbody>
</table>

The leaderboard ranks systems using unstemmmed, untokenized ROUGE-1 F-score by default, in order to fully reflect the difficulty of the summarization task, account for generated summary length, and measure performance most comparably across systems. Explore other stemming, tokenization, and ROUGE score variants above. HUM is the composite score of the NEWSROOM human evaluation task.

Click system names in the table to explore their performance in depth and read example summaries.

• 1.3 million articles and summaries.
• From 38 publications across 20 years.
• Wide range of summarization styles.

• Useful for training and evaluation.
• Allows comprehensive evaluation across publications, dates, and style.
• Combined with multi-dimensional human evaluation task.

• Data + leaderboard at summari.es
  (ask me for a flash drive)