Bloomberg

EntSUM: A Dataset for Entity-Centric Summarization

Engineering

Mounica Maddela* mmaddela3@gatech.edu, Mayank Kulkarni* mkulkarni24@bloomberg.net, Daniel Preoţiuc-Pietro dpreotiucpie@bloomberg.net

1. Entity-Centric Summarization

Given an article and an entity in the article, summarize the

salient content about the entity.



Input Article:

Department of Justice:
The Department of Justice shares
responsibility for disaster law enforcement
with the Homeland Security. The Justice
Department, which is in charge of the
F.B.I. and has long standing relations with
the local and state police, can rapidly
deploy law enforcement officers to
maintain public order.

Federal Emergency Medical Response:

The Department of Health would take back the disaster medical teams from FEMA. Housing and Urban Development would find temporary housing for victims, a duty handled by FEMA.

Department of Housing and Urban Development: The Department of Housing and Urban

Dovelopment would find temporary to the transfer and a species with board in the transfer and the second crisis.

Dovelopment would find temporary to the transfer and the second crisis with the second crisis. The second crisis was a second to EEE and the second crisis was a second to EEE and the second crisis was a second to EEE and the second crisis was a second to EEE and the second crisis was a second to EEE and the second crisis was a second to EEE and the second crisis was a s

3. Our Methods

Adapted SOTA document summarization approaches.

- $\bullet \ \ \textbf{GSum} \ \text{is an } \textbf{abstractive} \ \text{summarization approach}.$
- Consists of two encoders: input and guidance signal.
- Entity name / sentences as the guidance signal



Heuristics to obtain the reference summaries while training.

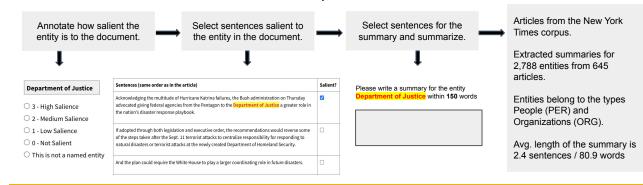
- Sentences in the overall reference summary containing the entity for up to 3 sentences.
- If no such sentences, take lead 3 sentences that mention the entity in the article.



- SentBERTSum is an extractive summarization approach.
- BERT-based classifier to select summary sentences.
- Input is a list of sentences containing the entity.

2. Construction of EntSUM

Created EntSUM, the first human-annotated benchmark for entity-centric and also controllable summarization.



4. Baselines

- Lead3 overall takes the first three sentences of the article.
- Lead3 entity takes the first three sentences containing the entity including coreferences.
- CTRLSum (He et. al 2020) and CNN enc dec (Fan et. al 2016) are existing abstractive approaches that prepend entity tokens to the input article.

All the methods and baselines are trained on The New York Times corpus and tested on the EntSUM dataset.

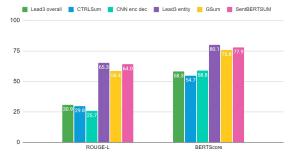
Code / Data

Takeaways

aithub.com/bloomberg/entsum

- Proposed new task called entity-centric summarization.
- Created new benchmark dataset.
- Adapted document summarization methods.

5. Results



- Lead3 overall performs poorly
- Existing entity prepending methods (CTRLSum and CNN) perform poorly.
- · Lead3 entity is a strong baseline
- GSum and SentBERTSum perform comparably to Lead 3 entity.
- · EntSUM is a challenging data set

* equal contribution