A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal

Demian Gholipour Ghalandari*+, Chris Hokamp+, Nghia The Pham+,
John Glover+, Georgiana Ifrim*

*School of Computer Science
University College Dublin, Ireland

+AYLIEN

Multi-Document Summarization for News

Collection of news articles

A meteorite strikes near the Cuban town of Viñales, in the western province of Pinar del Río, after sightings of a fireball over the Florida Keys. The last confirmed meteorite to hit Cuba was in 1994.

Summary

Previous Datasets for (News) Multi-Document Summarization

- Too small, no training possible (50-100 clusters in DUC, TAC)
- Small document clusters (2-3 docs in MultiNews)
- Unrealistic: clean, hand-picked input

Not representative for use cases with large & noisy clusters:

- Summaries for news clustering applications
- Summaries for search results
- Event summaries for timeline generation

Building the WCEP Dataset: Ground-truth & Source Articles

The Wikipedia Current Events Portal (WCEP)

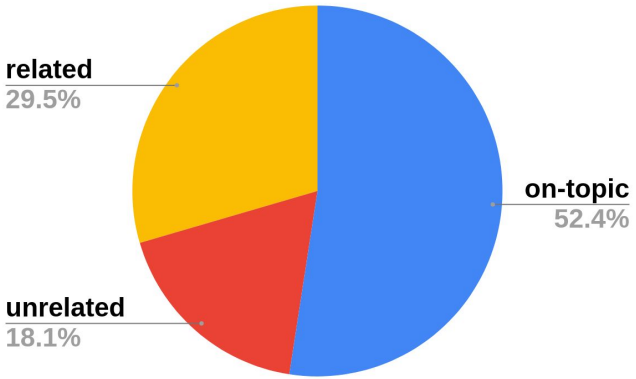
Not very “multi-document” so far

- Only 1.2 source articles on average
- Can we find additional, related news articles?

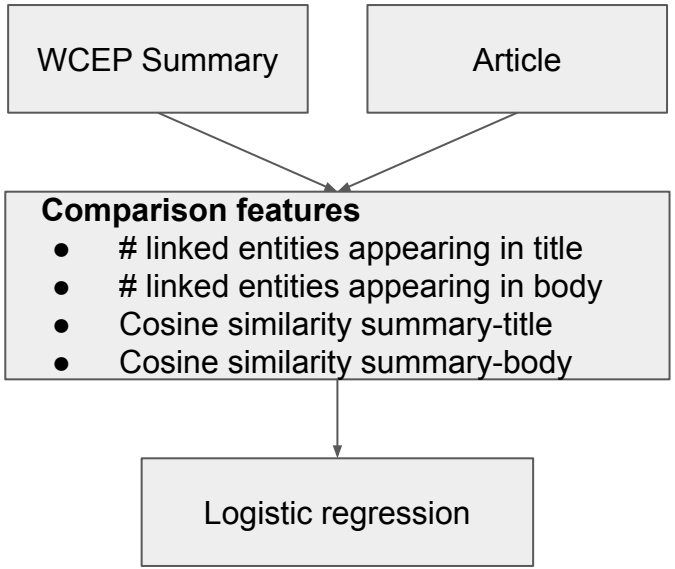
Obtaining Supplementary Articles from Common Crawl

For each WCEP summary, we add articles from Common Crawl matching these criteria

- Written in English
- Published within ± 1 day of the event date
- Classified as related to summary with prob 0.9



How many articles added from CommonCrawl are noisy?



Classifying article from Common Crawl to be a source of WCEP summary

Example

WCEP summary:
Reiwa (令和) is revealed as the new Japanese Era name set to start on May 1 upon Crown Prince Naruhito's accession to the Chrysanthemum Throne as the 126th Emperor of Japan.

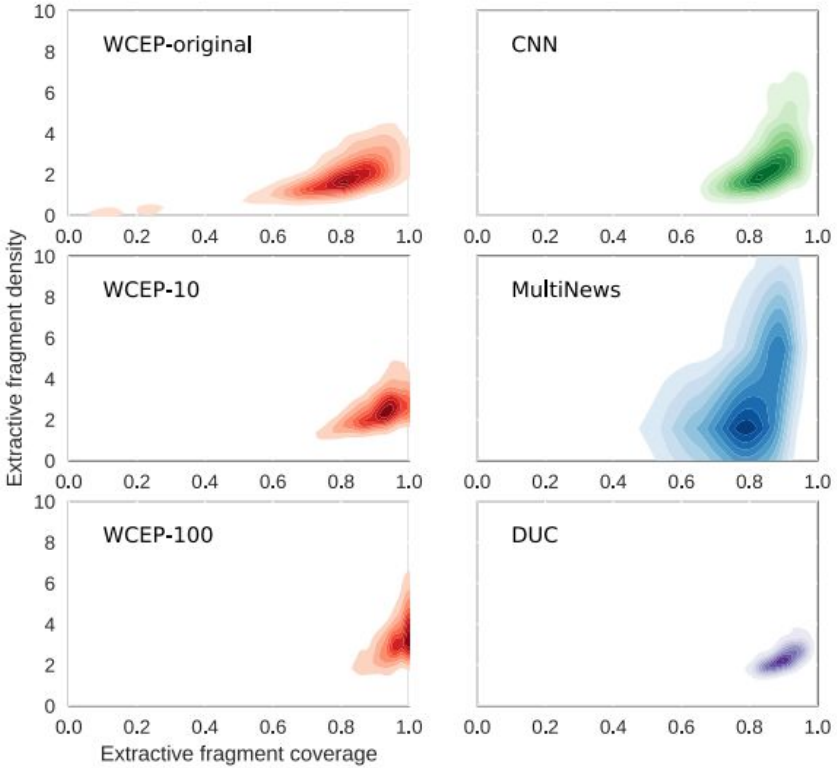
WCEP source:
Reiwa: Japan reveals name of new era ahead of Emperor's abdication

Similar articles from CommonCrawl (397 in total):

- Japan enters new era
- The Latest: Name for new era of Naruhito to be 'Reiwa'
- Nostalgia, excitement as Japan learns new imperial era name
- Reiwa: Japan reveals name of new era ahead of Emperor's abdication
- Defining Japan in just one word
- Japan gov't says era name translates as 'beautiful harmony'
- [...]

Dataset Statistics

- 10,200 clusters
- Average cluster size (#articles)
  - Only WCEP sources: **1.2**
  - WCEP + Common Crawl: **235**
- Average summary length
  - 32 words, 1.4 sentences
- “Extractiveness”
  - Summary tokens tend to be completely covered in large clusters
  - Copies of long sequences not common compared to MultiNews



Experiments

- Cluster size capped at 100 for experiments
- 40 token limit for the summary length
- More supplementary articles -> better
- Supervised extractive methods > unsupervised

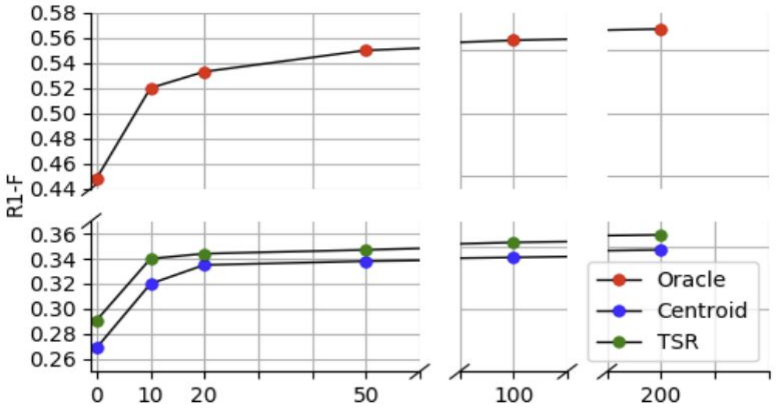


Figure 2: ROUGE-1 F1-scores for different numbers of supplementary articles from Common Crawl.

F-score			
Method	R1	R2	RL
ORACLE (MULTI)	0.558	0.29	0.4
ORACLE (SINGLE)	0.539	0.283	0.401
LEAD ORACLE	0.329	0.131	0.233
RANDOM LEAD	0.276	0.091	0.206
RANDOM	0.181	0.03	0.128
TEXTRANK	0.341	0.131	0.25
CENTROID	0.341	0.133	0.251
SUBMODULAR	0.344	0.131	0.25
TSR	**0.353**	**0.137**	**0.257**
BERTREG	0.35	0.135	0.255
SUBMODULAR+ABS	0.306	0.101	0.214
Recall			
Method	R1	R2	RL
ORACLE (MULTI)	0.645	0.331	0.458
ORACLE (SINGLE)	0.58	0.304	0.431
LEAD ORACLE	0.525	0.217	0.372
RANDOM LEAD	0.281	0.094	0.211
RANDOM	0.203	0.034	0.145
TEXTRANK	0.387	0.152	0.287
CENTROID	0.388	0.154	0.29
SUBMODULAR	0.393	0.15	0.289
TSR	**0.408**	**0.161**	**0.301**
BERTREG	0.407	0.16	**0.301**
SUBMODULAR+ABS	0.363	0.123	0.258

Unsupervised

Fully/part supervised

Table 5: Evaluation results on test set.

Conclusions & Future Work

- New large-scale MDS dataset with large clusters
- What is the dataset useful for?
  - Building models for short news event summaries
  - Researching scalable multi-input models
  - Robust evaluation of MDS methods due to large dataset size
- Pretraining & fine-tuning likely needed for abstractive summarisation
- Code for obtaining the dataset is available at: https://github.com/complementizer/wcep-mds-dataset