
[Data Science](#)[Natural Language Processing](#)[NLP Papers Summary](#)

Day 151: NLP Papers Summary – A Large-Scale Multi-Document Summarization Dataset From The Wikipedia Current Events Portal

By Ryan 30th May 2020 No Comments

Objective and Contribution

Presented a large dataset for multi-document summarisation (MDS) built from Wikipedia Current Events Portal (WCEP) that contains 10200 document clusters and each document cluster has 235 articles on average. Our dataset uses concise and neutral human-written summaries of news events, with links to external source articles. We extended the number of source articles by looking at related articles of the source articles in Common Crawl  re as

shown in the example below. We applied various supervised and unsupervised MDS models to establish baseline results for future research.

Human-written summary Emperor Akihito abdicates the Chrysanthemum Throne in favor of his elder son, Crown Prince Naruhito. He is the first Emperor to abdicate in over two hundred years, since Emperor Kōkaku in 1817.
Headlines of source articles (WCEP) <ul style="list-style-type: none">Defining the Heisei Era: Just how peaceful were the past 30 years?As a New Emperor Is Enthroned in Japan, His Wife Won't Be Allowed to Watch
Sample Headlines from Common Crawl <ul style="list-style-type: none">Japanese Emperor Akihito to abdicate after three decades on throneJapan's Emperor Akihito says he is abdicating as of Tuesday at a ceremony, in his final official address to his peopleAkihito begins abdication rituals as Japan marks end of era

Table 1: Example event summary and linked source articles from the Wikipedia Current Events Portal, and additional extracted articles from Common Crawl.

Dataset Construction

The dataset construction has three steps:

1. *Wikipedia Current Events Portal*. WCEP lists out daily news events whereby each news event has a human summary with at least one external news articles
2. *Obtaining Articles Linked on WCEP*. Each individual events contain a list of URLs to external source articles which we extracted all of it
3. *Additional Source Articles*. Additionally, we extended the input articles for each of the ground-truth summaries by searching for similar articles in the Common Crawl News dataset. We do this by training a simple logistic regression classifier to decide whether to assign an article to a summary. Our logistic regression has four different features as shown below:



tf-idf similarity between title and summary
tf-idf similarity between body and summary
No. entities from summary appearing in title
No. linked entities from summary appearing in body

Table 2: Features used in the article-summary binary classifier.

Overall, our final dataset consists of a ground-truth summary and a cluster of original source articles and related articles. The figure below showcase the summary statistics of the WCEP dataset and the statistics for individual clusters.

	TRAIN	VAL	TEST	TOTAL
# clusters	8,158	1,020	1,022	10,200
# articles (WCEP-total)	1.67m	339k	373k	2.39m
# articles (WCEP-100)	494k	78k	78k	650k
period start	2016-8-25	2019-1-6	2019-5-8	-
period end	2019-1-5	2019-5-7	2019-8-20	-

Table 3: Size overview of the WCEP dataset.

	MIN	MAX	MEAN	MEDIAN
# articles (WCEP-total)	1	8411	234.5	78
# articles (WCEP-100)	1	100	63.7	78
# WCEP articles	1	5	1.2	1
# summary words	4	141	32	29
# summary sents	1	7	1.4	1

Table 4: Stats for individual clusters in WCEP dataset.

QUALITY OF ADDITIONAL ARTICLES

To ensure that our additional articles from Common Crawl are related to our source articles, we manually annotated 350 additional articles. We compare the article title with the first three sentences of the assigned summary and label the following:

1. *On-topic*. When the article focuses on the event described in the summary

2. *Related*. When the article mentions the event but focuses on other things

3. *Unrelated*. When the article has no mention of the event



We have 52% on-topic and 30% related additional articles from the 350 articles.

MEASURING EXTRACTIVENESS OF OUR SUMMARIES

Here, we aim to measure how extractive our summaries are by using the coverage and density metrics. Coverage measures the number of words from the summary that's extracted from all the articles in a cluster whereas the density measures how well a summary can be described as a series of extractions. The results of the two measures are shown below. The WCEP dataset shows high coverage if articles are included from Common Crawl. This means that copy mechanisms would be useful for generating summaries.

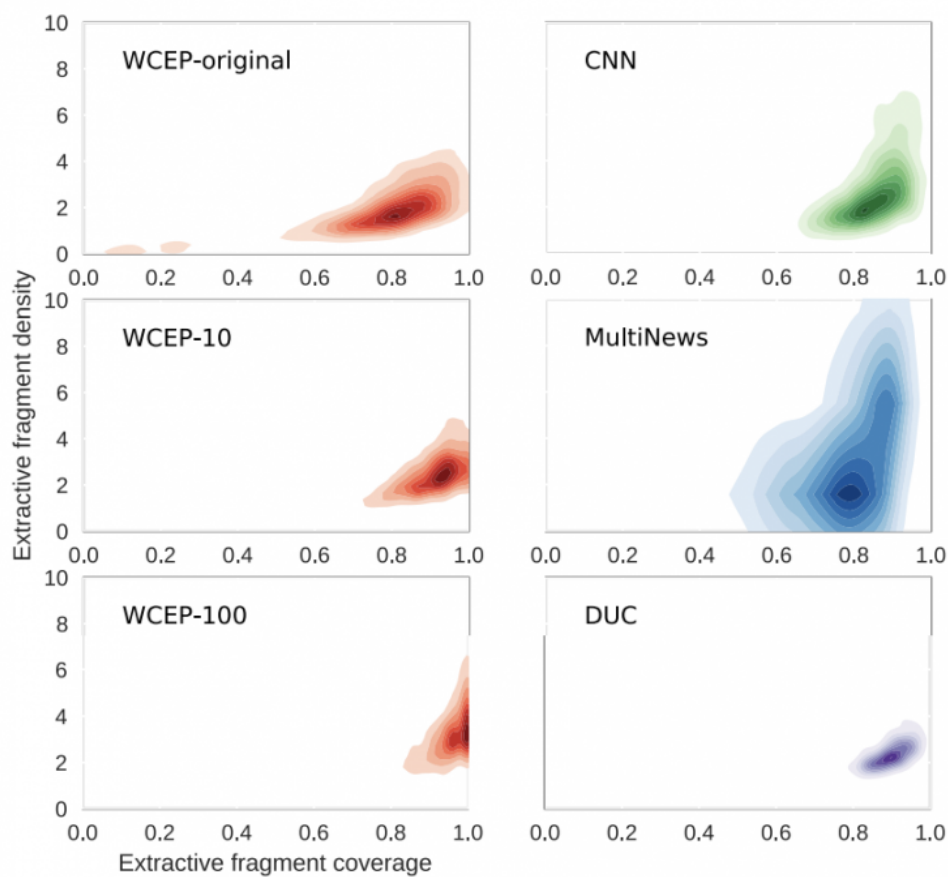



Figure 1: Coverage and density on different summarization datasets.

Experiments and Results

We only considered 100 articles per cluster (WCEP-100) due to scalability and th  that performance starts to plateau after 100 articles. Our evaluation metrics are the F1 score and

Recall score of ROUGE-1, ROUGE-2, and ROUGE-L. We considered different oracle and baseline models to allow us to a) measure the upper bound of our performance and b) evaluate our dataset using common SOTA models.

RESULTS

The table below showcases the results. As shown, there's still a wide margin between the oracle results and the best performing model, indicating more research needs to be done. In addition, the supervised methods seem to outperform the unsupervised methods but only by a small margin. The high single document oracle result tells us that it is important to select relevant articles before summarisation. Lastly, the dataset doesn't support lead summaries like other summarisation datasets as demonstrated by the low performance of RANDOM LEAD.



Conclusion and Future Work

Potential work involves how to scale deep learning models to the huge MDS corpus and how to close the gap between existing methods and the oracle results.

Source: <https://arxiv.org/pdf/2005.10070.pdf>



Ryan

Data Scientist



Previous Post

Day 150: NLP
Papers Summary -
Will-They-Won't-
They: A Very Large
Dataset for Stance
Detection on
Twitter

Next Post

Day 152: NLP
Papers Summary -
OPINIONDIGEST:
A Simple
Framework for
Opinion
Summarization





[Data Science](#) [Natural Language Processing](#) [NLP Papers Summary](#)

Day 365: NLP Papers Summary – A Survey on Knowledge Graph Embedding

Ryan

30th December 2020





[Data Science](#) [Implementation](#) [Natural Language Processing](#)

Day 364: Ryan's PhD Journey – OpenKE-PyTorch Library Analysis + code snippets for 11 KE models

Ryan

29th December 2020





[Data Science](#) [Ryan's PhD Journey](#)

Day 363: Ryan's PhD Journey – Literature Review – Knowledge Acquisition – 1st Passes XIV

Ryan

28th December 2020

