Tixierae /
OrangeSum

Code Issues Pull requests Actions Projects Security Insights

The French summarization dataset introduced in "BARThez: a Skilled Pretrained French Sequence-to-Sequence Model".

arxiv.org/pdf/2010.12321.pdf

CC-BY-SA-4.0 license

22 stars 2 forks 3 watching 1 Branch 0 Tags Activity

Public repository

main 1 Branch 0 Tags Go to file t Go to file + Add file Code ···

Tixierae Update README.md 3 years ago

data	Add summarization splits compressed	4 years ago
figures	more readable	4 years ago
plots	first commit	4 years ago
splits_heading_as_summary	first commit	4 years ago
splits_title_as_summary	first commit	4 years ago
summaries	update	4 years ago
LICENSE.txt	Added license	3 years ago
README.md	Update README.md	3 years ago
Visualization.ipynb	first commit	4 years ago
compute_overlap.py	first commit	4 years ago
filter_split.py	first commit	4 years ago
generate_tables.R	first commit	4 years ago
get_urls.py	first commit	4 years ago
list_of_duplicates.txt	first commit	4 years ago
parse_urls.py	first commit	4 years ago
random_examples.json	first commit	4 years ago
scrape_urls.py	first commit	4 years ago
statistics.py	first commit	4 years ago
	first commit	4 years ago

summary_statistics.R

README CC-BY-SA-4.0 license

OrangeSum Dataset

What is this repo for?

This repository provides the French summarization dataset introduced in the paper BARThez: a Skilled Pretrained French Sequence-to-Sequence Model (Kamal Eddine, Tixier, and Vazirgiannis, 2020), with the train, development, and test splits used in the paper. It also provides the code that was used to build the dataset, and the test set summaries generated by the BARThez, mBART, mBARThez, and CamemBERT2CamemBERT models, to make cross comparison with future work as easy as possible (in `./summaries/`).

Note: this repository is dedicated to the OrangeSum dataset. The main repository of the paper is https://github.com/moussaKam/BARThez.

OrangeSum

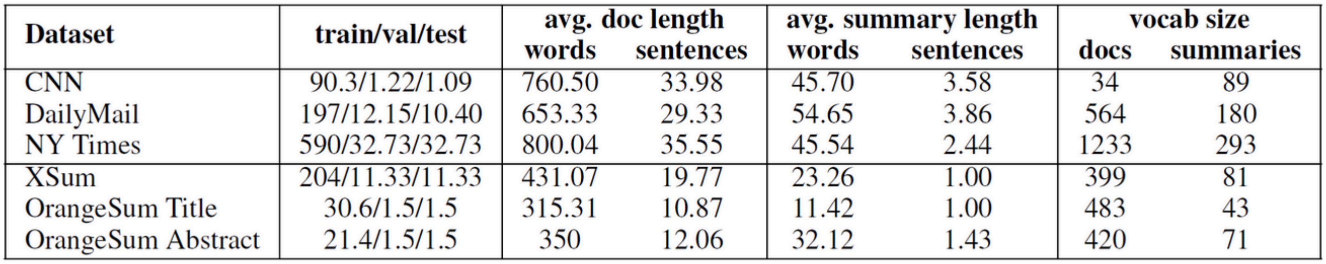
The OrangeSum dataset was inspired by the XSum dataset. It was created by scraping the "Orange Actu" website: https://actu.orange.fr/. Orange S.A. is a large French multinational telecommunications corporation, with 266M customers worldwide. Scraped pages cover almost a decade from Feb 2011 to Sep 2020. They belong to five main categories: France, world, politics, automotive, and society. The society category is itself divided into 8 subcategories: health, environment, people, culture, media, high-tech, unusual ("insolite" in French), and miscellaneous.

Each article featured a single-sentence title as well as a very brief abstract, both professionally written by the author of the article. These two fields were extracted from each page, thus creating two summarization tasks: **OrangeSum Title** and **OrangeSum Abstract**.

As a post-processing step, we removed all empty articles, and articles whose titles were shorter than 5 words. For OrangeSum Abstract, we removed the top 10% articles in terms of proportion of novel unigrams in the abstracts, as we observed that such abstracts tended to be introductions rather than real abstracts. This corresponded to a threshold of 57% novel unigrams.

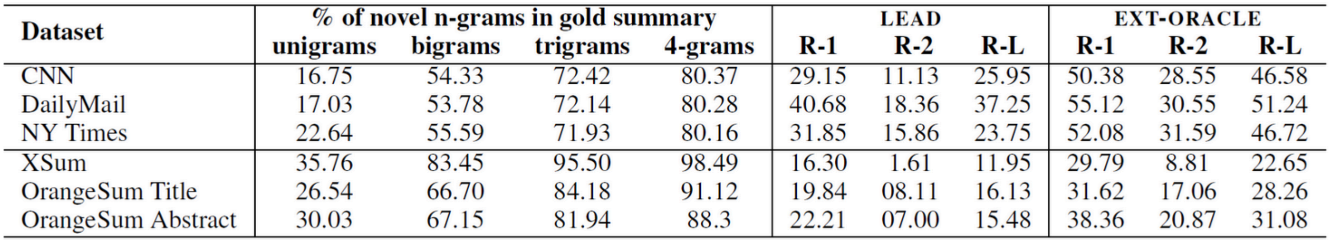
For both OrangeSum Title and OrangeSum Abstract, we set aside 1500 pairs for testing, 1500 for validation, and used all the remaining ones for training.

In the table below, Sizes (column 2) are given in thousands of documents, document and summary lengths are in words, and vocab sizes are in thousands of tokens.



Dataset	train/val/test	avg. doc length		avg. summary length		vocab size	
		words	sentences	words	sentences	docs	summaries
CNN	90.3/1.22/1.09	760.50	33.98	45.70	3.58	34	89
DailyMail	197/12.15/10.40	653.33	29.33	54.65	3.86	564	180
NY Times	590/32.73/32.73	800.04	35.55	45.54	2.44	1233	293
XSum	204/11.33/11.33	431.07	19.77	23.26	1.00	399	81
OrangeSum Title	30.6/1.5/1.5	315.31	10.87	11.42	1.00	483	43
OrangeSum Abstract	21.4/1.5/1.5	350	12.06	32.12	1.43	420	71

In the table below, it can be observed that OrangeSum offers approximately the same degree of abstractivity as XSum, and that both of them are more abstractive than traditional summarization datasets.



Dataset	% of novel n-grams in gold summary				LEAD			EXT-ORACLE		
	unigrams	bigrams	trigrams	4-grams	R-1	R-2	R-L	R-1	R-2	R-L
CNN	16.75	54.33	72.42	80.37	29.15	11.13	25.95	50.38	28.55	46.58
DailyMail	17.03	53.78	72.14	80.28	40.68	18.36	37.25	55.12	30.55	51.24
NY Times	22.64	55.59	71.93	80.16	31.85	15.86	23.75	52.08	31.59	46.72
XSum	35.76	83.45	95.50	98.49	16.30	1.61	11.95	29.79	8.81	22.65
OrangeSum Title	26.54	66.70	84.18	91.12	19.84	08.11	16.13	31.62	17.06	28.26
OrangeSum Abstract	30.03	67.15	81.94	88.3	22.21	07.00	15.48	38.36	20.87	31.08

Steps to create the dataset

Starting from an empty directory structure, run the following scripts, in that order.

1. `get_urls.py`
2. `scrape_urls.py`
3. `parse_urls.py`
4. `compute_overlap.py`
5. `filter_split.py`

Notes

1. Some of the articles that were scraped might not still be online. The raw HTML files were saved and are released here, though.
2. Sometimes, "heading" is used in the code and the repository. It corresponds to the Abstract task in the paper.
3. The dataset was augmented by running a second round of scraping about two months after the initial one, to collect new articles. In this process, a line `========` was appended at the end of the `urls.txt` file. The indexes of the new documents start from the index of the following line (in `urls.txt`). These indexes were passed to the `scape_one_url()` function that writes the documents, but the new URLs were appended directly at the end of the `urls_final.txt` file. This created an index gap as `urls_final.txt` had not the same number of lines as `urls.txt` at the beginning of the process. **So, to sum up, there is a perfect mapping between the line numbers of the URLs and the final `.json` files from 0 to 31134. Then, one needs to add 236, i.e., URL index + 236 = `.json` index.**

Cite

If you use our code or dataset, please cite:

BibTex

```
@article{eddine2020barthez,
  title={BARThez: a Skilled Pretrained French Sequence-to-Sequence Model},
  author={Eddine, Moussa Kamal and Tixier, Antoine J-P and Vazirgiannis, Michalis},
  journal={arXiv preprint arXiv:2010.12321},
  year={2020}
}
```

MLA

> Eddine, Moussa Kamal, Antoine J-P. Tixier, and Michalis Vazirgiannis. "BARThez: a Skilled Pretrained French Sequence-to-Sequence Model." arXiv preprint arXiv:2010.12321 (2020).

Releases

No releases published

Packages

No packages published

Contributors 2

Tixierae Antoine J.-P. Tixier

moussaKam Moussa Kamal Eddine

Languages

● **Jupyter Notebook** 91.9% ● **Python** 6.6% ● **Other** 1.5%