PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization - Wen Xiao, Iz Beltagy, Giuseppe Carenini, Arman Cohan (UBC + Allen AI), ACL 2022

Presentation Author
Mathieu Ravaut

Nanyang Technological University

Table of content

Introduction

Current state-of-the-art (single-stage) abstractive document summarization systems rely on pre-trained encoder-decoder models :

- With a general generation pre-training objective :
  - T5 (multiple text-to-text tasks) [14]
  - BART (sequence denoising) [10]
  - ProphetNet (predicting multiple future n-grams) [13]
- With a pre-training objective tailored to summarization :
  - PEGASUS [16]
  - TED (quite similar to PEGASUS) [15]

Introduction

Remember the PEGASUS key idea : **Gap Sentences Generation**.

- For each sentence in the document, compute its ROUGE-1 against the *rest* of the document. This is a proxy method to find *salient* sentences.
- Take out the 3 sentences with the highest ROUGE-1 as a pseudo-summary.
- Replace them with [MASK1].
- Pre-train the model to predict these 3 sentences *based on the rest of the document.*

Introduction

What about **multi-document** summarization ? What are the options ?

It's not so straightforward to extend PEGASUS :

- Option 1 : take salient sentences from each document with regards to the rest of this document.
- Option 2 : take salient sentences from each document with regards to the rest of this document and all other documents.

PEGASUS uses option 2 for Multi-News [6].

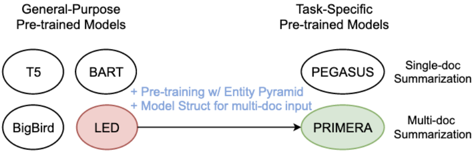
Introduction

This paper PRIMERA makes the following assumption :

Cluster of multiple documents typically include redundant information. The GSG salient sentences selection method favors an exact match between sentences (due to ROUGE-1), which in the multi-document case, will miss out a lot of representative information.

Introduction

PRIMERA proposes another method to select the salient sentences for the GSG objective based on **entities**.



This results in a new pre-training objective tailored for multi-document summarization.
Other than sentence selection, the rest of the algorithm follows PEGASUS.

Model : intuition

The PRIMERA salient sentences identification strategy is inspired by the **Pyramid Evaluation framework** :

- Framework for evaluating summaries with multiple human written references.
- Human annotators chunk reference summaries in Summary Content Unit (SCUs) (words or phrases).
- SCUs receive a score proportional to the number of reference summaries containing them.
- The candidate summary score is the normalized mean of the SCUs it contains.

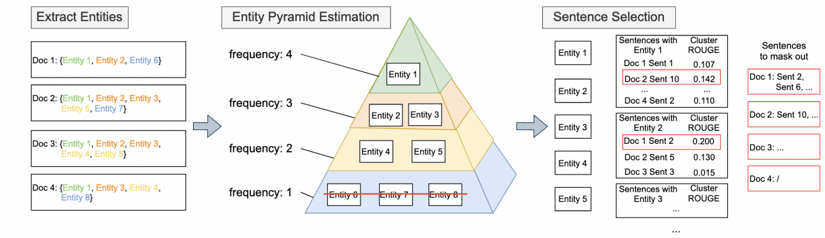
Model : salient sentence selection

The PRIMERA salient sentences identification strategy is the following :

- Find all entities in the cluster with Spacy.
- Calculate the saliency of entities based on their frequency in the cluster.
- Sort entities by decreasing frequency, remove the ones with frequency of just 1.
- Until m sentences have been found :
  - Start from the most salient entities to the least salient ones.
  - Take all sentences in the cluster containing the entity.
  - From the subset above, take the sentence maximizing overlap with the *other documents than the one it appears in.*

Model : salient sentence selection

Illustation of the previous process :



Model : encoder-decoder

PRIMERA uses as backbone encoder-decoder model the Longformer Encoder-Decoder (LED) [1] :

- Combination of local and global attention.
- Can scale to input length 4096 and output length 1024.
- Use sliding window size 512 for local attention.

Documents are concatenated and truncated to the max input length divided by the number of documents.

Experiments : datasets

Pre-training on Newshead [8].
Fine-tuning on : Multi-News [6], Multi-Xscience [11], Wikisum [3], WCEP-10 [7], DUC2004 [4], arXiv [2].

Dataset	#Examples	#Doc/C	Len_{src}	Len_{summ}
Newshead (2020)	360K	3.5	1734	-
Multi-News (2019)	56K	2.8	1793	217
Multi-Xscience (2020)	40K	4.4	700	105
Wikisum* (2018)	1.5M	40	2238	113
WCEP-10 (2020)	10K	9.1	3866	28
DUC2004 (2005)	50	10	5882	115
arXiv (2018)	214K	5.5	6021	272

Experiments : zero-shot

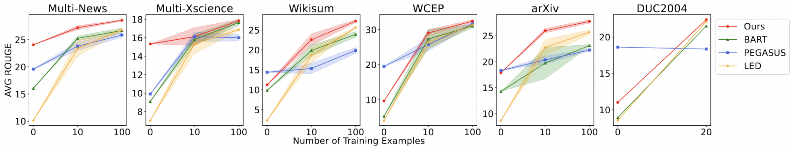
Zero-shot results :

Models	Multi-News(256)			Multi-XSci(128)			WCEP(50)			WikiSum(128)			arXiv(300)			DUC2004 (128)		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
PEGASUS★(Zhang et al., 2020)	36.5	10.5	18.7	-	-	-	-	-	-	-	-	-	28.1	6.6	17.7	-	-	-
PEGASUS (our run)	32.0	10.1	16.7	27.6	**4.6**	15.3	**33.2**	**12.7**	**23.8**	24.6	5.5	15.0	29.5	7.9	17.1	32.7	**7.4**	17.6
BART (our run)	27.3	6.2	15.1	18.9	2.6	12.3	20.2	5.7	15.3	21.6	5.5	15.0	29.2	7.5	16.9	24.1	4.0	15.3
LED (our run)	17.3	3.7	10.4	14.6	1.9	9.9	18.8	5.4	14.7	10.5	2.4	8.6	15.0	3.1	10.8	16.6	3.0	12.0
PRIMERA (our model)	**42.0**	**13.6**	**20.8**	**29.1**	4.6	**15.7**	28.0	10.3	20.9	**28.0**	**8.0**	**18.0**	**34.6**	**9.4**	**18.3**	**35.1**	7.2	**17.9**

+2-3 points compared to PEGASUS, except on WCEP.

Experiments : few-shot

Few-shot (10, 100) results, averaged over 5 runs :



Best method in all datasets.

Experiments : full supervised

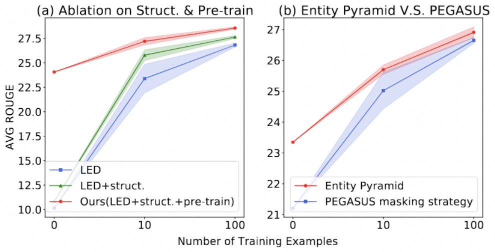
Full-data results :

Datasets	Previous SOTA			PRIMERA		
	R-1	R-2	R-L	R-1	R-2	R-L
Multi-News	49.2	19.6	24.5	**49.9**	**21.1**	**25.9**
Multi-XScience	**33.9**	6.8	**18.2**	31.9	**7.4**	18.0
WCEP	35.4	15.1	25.6	**46.1**	**25.2**	**37.9**
arXiv	46.6	19.6	41.8	**47.6**	**20.8**	**42.6**

New SOTA on Multi-News (+0.7 R-1), WCEP (+10.7 R-1) and arXiv (+1.0 R-1).

Experiments : ablation

Ablation study :



The new pre-training objective contributes a lot.

Experiments : qualitative evaluation

Human evaluation : Pyramid scores on DUC-2007 and TAC-2008.

Model	DUC2007(20)				TAC2008(20)			
	S_r	R	P	F	S_r	R	P	F
PEGASUS	6.0	2.5	2.4	2.4	**8.7**	**9.1**	9.4	9.1
LED	9.6	3.9	4.0	3.8	6.9	7.1	**10.8**	8.4
PRIMERA	**12.5**	**5.1**	**5.0**	**5.0**	8.5	8.9	10.0	**9.3**

Table 4: Pyramid Evaluation results: Raw scores S_r, (R)ecall, (P)recision and (F)-1 score. For readability, Recall, Precision and F-1 scores are multiplied by 100.

PRIMERA is better on DUC-2007.

Experiments : qualitative evaluation

Human evaluation : fluency.

Model	DUC2007(20)			TAC2008(20)		
	Gram.	Ref.	Str.&Coh.	Gram.	Ref.	Str.&Coh.
PEGASUS	4.45	4.35	1.95	**4.40**	4.20	3.20
LED	4.35	4.50	3.20	3.10	3.80	2.55
PRIMERA	**4.70**	**4.65**	**3.70**	**4.40**	**4.45**	**4.10**

Table 5: The results of Fluency Evaluation on two datasets, in terms of the Grammaticality , Referential clarity and Structure & Coherence.

PRIMERA is better on both DUC-2007 and TAC-2008.

Conclusion

- New pre-training objective tailored to multi-document abstractive summarization.
- Echoes other work using entities in summarization : CTRLSum [9], GSum [5], FROST [12].
- Relying on the Longformer Encoder-Decoder.
- Convincing evaluation results :
    - 6 datasets
    - 3 domains : news, Wikipedia, science
    - 3 data volumes : zero-shot, few-shot, full supervised

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer : The long-document transformer. *arXiv preprint arXiv :2004.05150*, 2020.

[2] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv :1804.05685*, 2018.

[3] Nachshon Cohen, Oren Kalinsky, Yftah Ziser, and Alessandro Moschitti. Wikisum : Coherent summarization dataset for efficient human-evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, pages 212–219, 2021.

[4] Hoa Trang Dang. Overview of duc 2005. In *Proceedings of*

the document understanding conference, volume 2005, pages 1–12, 2005.

[5] Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. Gsum : A general framework for guided neural abstractive summarization. *arXiv preprint arXiv :2010.08014*, 2020.

[6] Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. Multi-news : A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv :1906.01749*, 2019.

[7] Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. A large-scale multi-document summarization dataset from the wikipedia current events portal. *arXiv preprint arXiv :2005.10070*, 2020.

[8] Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, You Wu, Cong Yu, Daniel Finnie, Hongkun Yu, Jiaqi Zhai, and Nicholas Zukoski. Generating representative headlines for news stories. In *Proceedings of The Web Conference 2020*, pages 1773–1784, 2020.

[9] Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. Ctrlsum : Towards generic controllable text summarization. *arXiv preprint arXiv :2012.04281*, 2020.

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv :1910.13461*, 2019.

[11] Yao Lu, Yue Dong, and Laurent Charlin. Multi-xscience :

A large-scale dataset for extreme multi-document summarization of scientific articles. *arXiv preprint arXiv :2010.14235*, 2020.

[12] Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9 :1475–1492, 2021.

[13] Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. Prophetnet : Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv :2001.04063*, 2020.

[14] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a

unified text-to-text transformer. *arXiv preprint arXiv :1910.10683*, 2019.

[15] Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. Ted : A pretrained unsupervised summarization model with theme modeling and denoising. *arXiv preprint arXiv :2001.00725*, 2020.

[16] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus : Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.