PRIMERA
Pyramid-based Masked Sentence Pre-training for Multi-document Summarization

Wen Xiao[1], Iz Beltagy[2], Giuseppe Carenini[1], Arman Cohan[2]

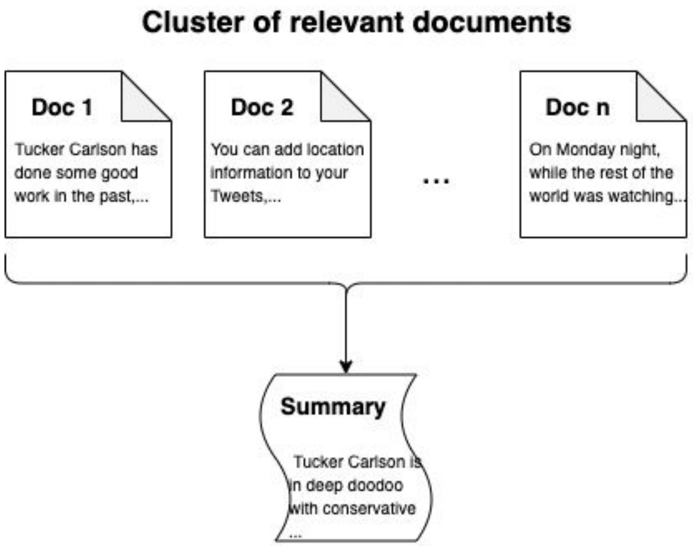
[1] University of British Columbia
[2] Allen Institute of Artificial Intelligence

What is Multi-Document Summarization (MDS)?

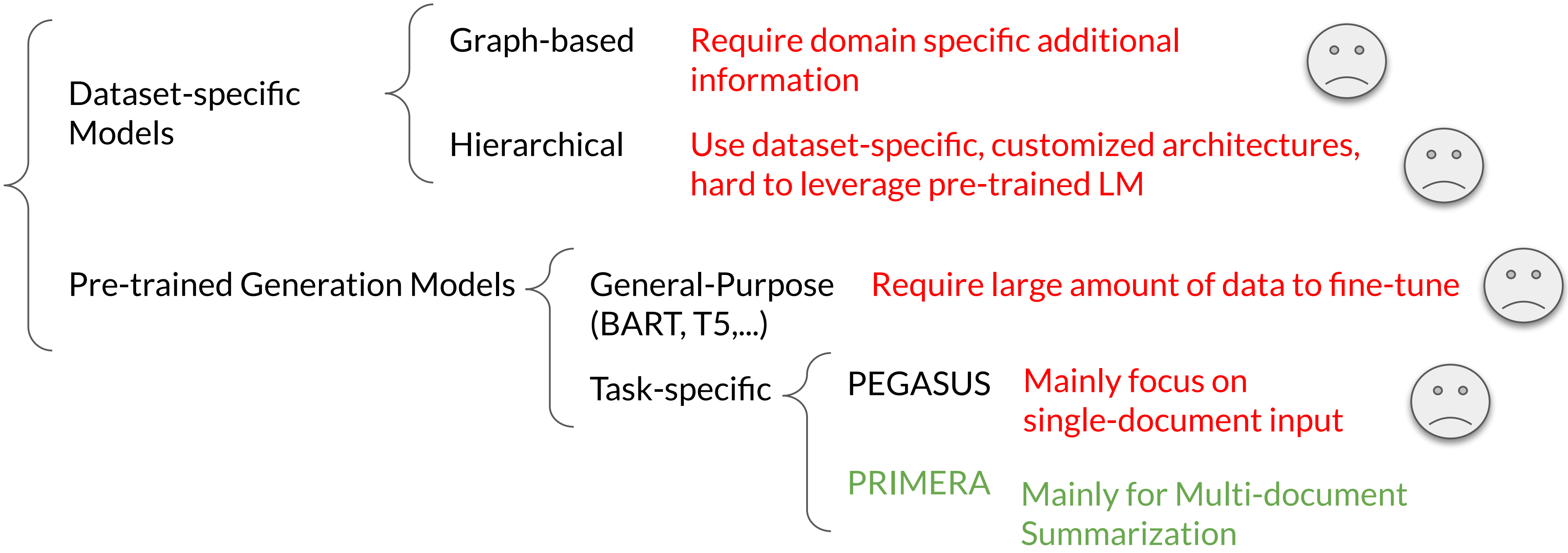
Task: Generate a summary given **a cluster of relevant documents,**

e.g.

- News articles
- Scientific papers


Cluster of relevant documents
Doc 1
Tucker Carlson has done some good work in the past,...
Doc 2
You can add location information to your Tweets,...
...
Doc n
On Monday night, while the rest of the world was watching...
Summary
Tucker Carlson is in deep doodoo with conservative
...

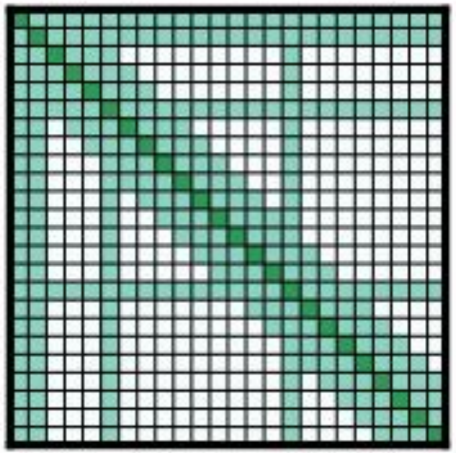
Previous Methods for MDS



- Dataset-specific Models
  - Graph-based — Require domain specific additional information ☹
  - Hierarchical — Use dataset-specific, customized architectures, hard to leverage pre-trained LM ☹
- Pre-trained Generation Models
  - General-Purpose (BART, T5,...) — Require large amount of data to fine-tune ☹
  - Task-specific
    - PEGASUS — Mainly focus on single-document input ☹
    - PRIMERA — Mainly for Multi-document Summarization

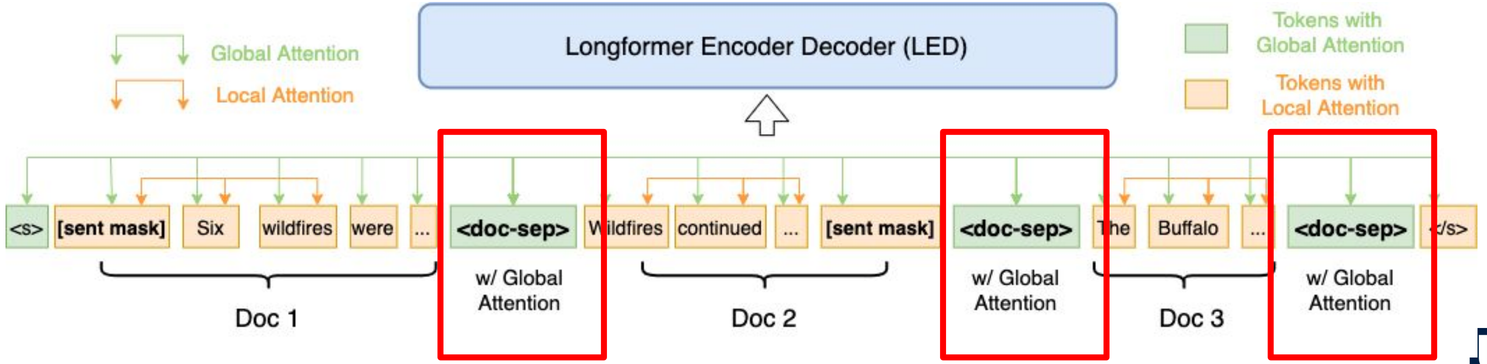
Overview of PRIMERA

- Architecture: Longformer Encoder Decoder (LED)
  - Global + Local Attention
  - Allows for long sequence inputs

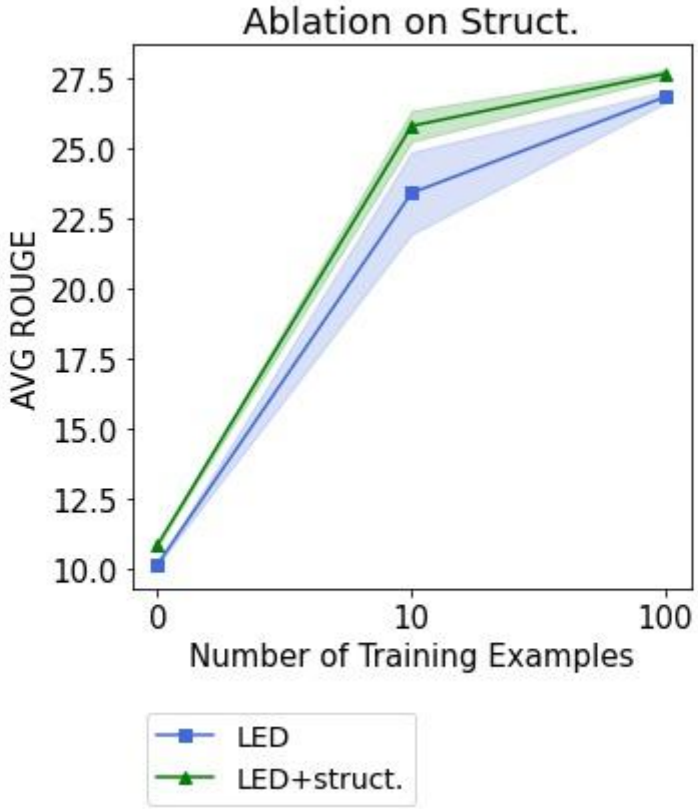


Overview of PRIMERA

- Architecture: Longformer Encoder Decoder (LED)
  - Global + Local Attention
  - Allows for long sequence input
- Input Structure:
  - documents separated with document separator(*<doc-sep>*)
  - Global Attention on *<doc-sep>*



Impact of the Proposed Input Structure



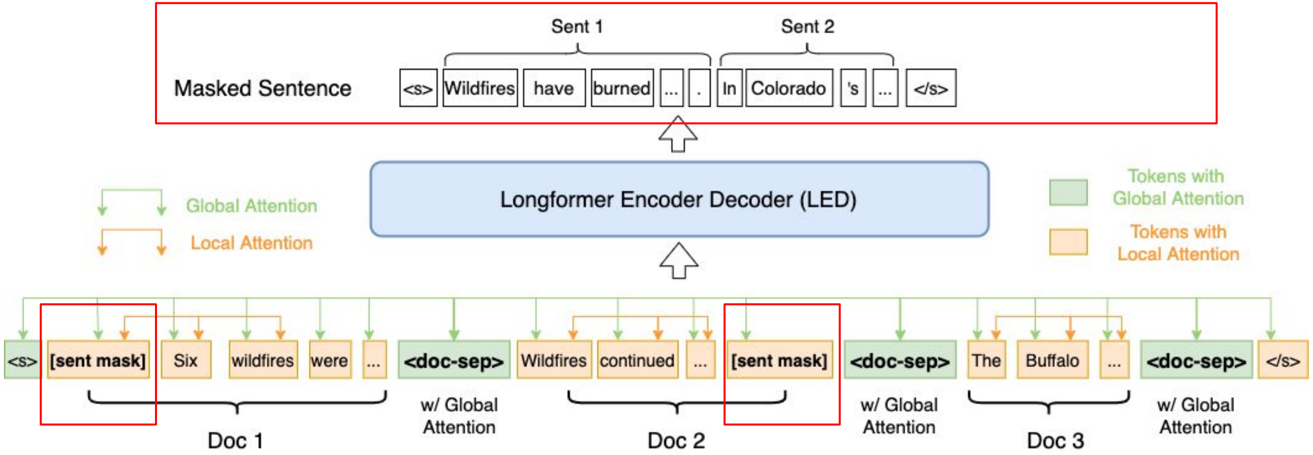
- Input structure improves the results

Overview of PRIMERA

- Architecture: Longformer (with local and global attention)
    - Global + Local Attention
    - Allows for long sequence input
- Input Structure:
    - documents separated with document separator(<*doc-sep*>)
    - Global Attention on <*doc-sep*>
- Pre-training:
    - **Goal:** Teach the model to identify and aggregate salient information across a “cluster” of related documents
    - **Multi-doc corpus:** Newshead (360k clusters, 3.5 doc/cluster on average)
    - **Objective**: Gap Sentence Generation (GSG)
    - **Novel Masking Strategy:** Entity Pyramid

Pre-training : Objective

- **Gap Sentence Generation** [Zhang et al., 2020]:
  - Select several SALIENT sentences from the input documents (as pseudo-summary)
  - Mask out the selected sentences
  - Generate them in order in the decoder



How to select SALIENT sentences?

Previous work for single document input (PEGASUS):

- Random
- Lead-K
- Principle Best
  - **Intuition**: select the **most central** sentences in the document
  - The score is defined as the ROUGE score between **each sentence** and **rest of the document**

$$\mathrm{Score}(s_i) = \mathrm{Rouge}(s_i, D / \{s_i\})$$

How to select SALIENT sentences?

- However, multi-document input tends to be more **redundant** than single document input.
- And such strategy would prefer **exact match between sentences**, resulting in selection of less representative information.

Example of the problem with vanilla SGS

Doc #1	Wildfires have burned across tens of thousands of acres of parched terrain in Colorado, spurring thousands of evacuations ..., residents have sought shelter in middle schools, and local officials fear tourists usually drawn to the region for the summer may not come.
Doc #2	... In Colorado's southwest, authorities have shuttered the San Juan National Forest in southwestern Colorado and residents of more than 2,000 homes were forced to evacuate. No homes had been destroyed ... "Under current conditions, one abandoned campfire or spark could cause a catastrophic wildfire, ..., with human life and property," said San Juan National Forest Fire Staff Officer Richard Bustamante...
Doc #3	The Buffalo Fire west of Denver is ... Several wildfires in Colorado have prompted thousands of home evacuations ... Nearly 1,400 homes have been evacuated in Summit County, Colorado, ... "Under current conditions, one abandoned campfire or spark could cause a catastrophic wildfire, ... , with human life and property," said Richard Bustamante, SJNF forest fire staff officer ...

New Masking Strategy: Entity Pyramid

Goal:

Select sentences that best represent the entire cluster of input documents

Inspired by Pyramid Evaluation

Pyramid Evaluation

with multiple refs

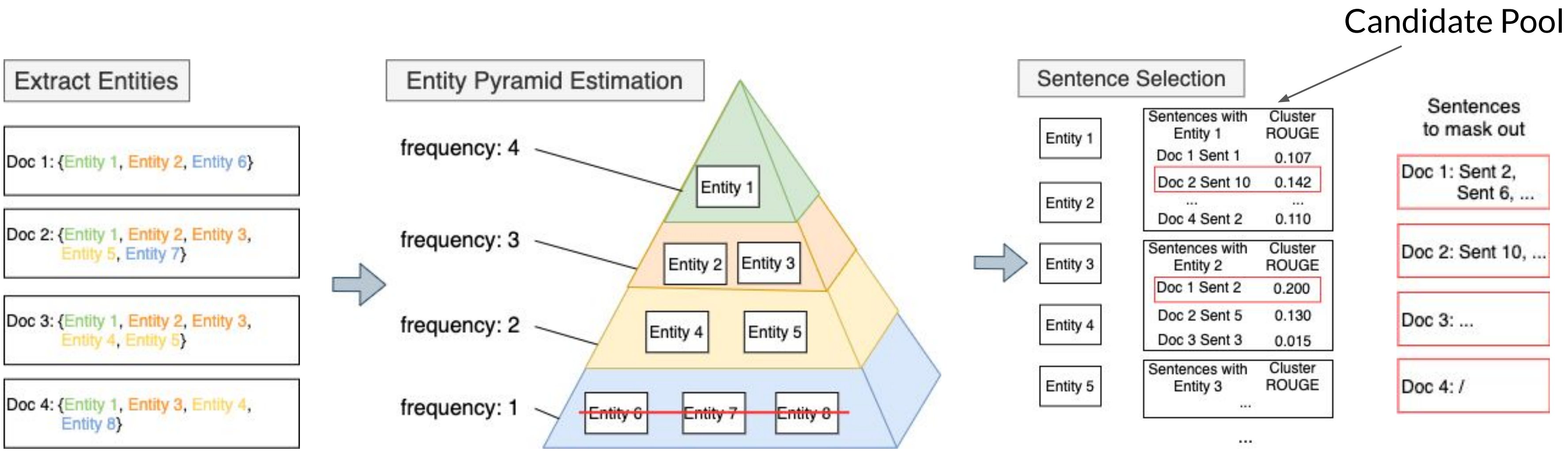
- The importance of information is quantified by the frequency of the gold references that include it.
- The more gold references include a fact, the more important it is.
- The facts are identified by human-labeled Summary Content Units (SCUs)

Entity Pyramid

with multiple docs

- The importance of information is quantified by the frequency of the documents that include it.
- The more documents include a fact, the more important it is.
- The facts are identified by Entities

New Masking Strategy: Entity Pyramid



Candidate Pool

Extract Entities

Doc 1: {Entity 1, Entity 2, Entity 6}

Doc 2: {Entity 1, Entity 2, Entity 3, Entity 5, Entity 7}

Doc 3: {Entity 1, Entity 2, Entity 3, Entity 4, Entity 5}

Doc 4: {Entity 1, Entity 3, Entity 4, Entity 8}

Entity Pyramid Estimation

frequency: 4 — Entity 1

frequency: 3 — Entity 2, Entity 3

frequency: 2 — Entity 4, Entity 5

frequency: 1 — ~~Entity 6~~, ~~Entity 7~~, ~~Entity 8~~

Sentence Selection

Entity 1

Entity 2

Entity 3

Entity 4

Entity 5

Sentences with Entity 1	Cluster ROUGE
Doc 1 Sent 1	0.107
Doc 2 Sent 10	0.142
...	...
Doc 4 Sent 2	0.110

Sentences with Entity 2	Cluster ROUGE
Doc 1 Sent 2	0.200
Doc 2 Sent 5	0.130
Doc 3 Sent 3	0.015

Sentences with Entity 3	Cluster ROUGE
...	

...

Sentences to mask out

Doc 1: Sent 2, Sent 6, ...

Doc 2: Sent 10, ...

Doc 3: ...

Doc 4: /

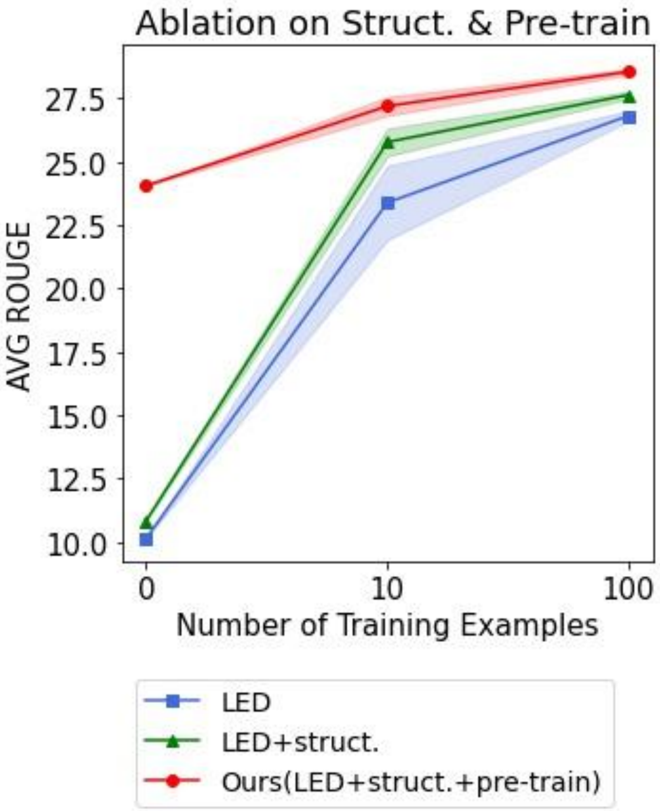
Cluster ROUGE:

$$Score(s_i) = \sum_{\{doc_j \in C, s_i \notin doc_j\}} \mathrm{ROUGE}(s_i, doc_j)$$

Example

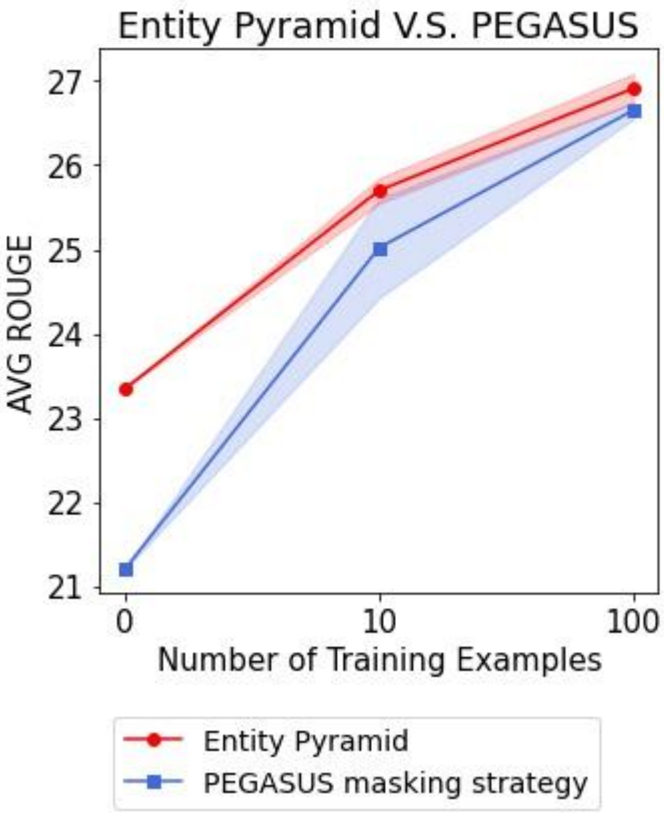
Doc #1	Wildfires have burned across tens of thousands of acres of parched terrain in Colorado, spurring thousands of evacuations (0.107) ..., residents have sought shelter in middle schools, and local officials fear tourists usually drawn to the region for the summer may not come.
Doc #2	... ***In Colorado's southwest, authorities have shuttered the San Juan National Forest in southwestern Colorado and residents of more than 2,000 homes were forced to evacuate. (0.187)*** No homes had been destroyed ... "Under current conditions, one abandoned campfire or spark could cause a catastrophic wildfire, ..., with human life and property," said San Juan National Forest Fire Staff Officer Richard Bustamante...
Doc #3	The Buffalo Fire west of Denver is ... Several wildfires in Colorado have prompted thousands of home evacuations (0.172)... Nearly 1,400 homes have been evacuated in Summit County, Colorado, (0.179)..... "Under current conditions, one abandoned campfire or spark could cause a catastrophic wildfire, ... , with human life and property," said Richard Bustamante, SJNF forest fire staff officer ...
Entity List	Colorado(3), Wildfires(3), 416(2), Tuesday(2), San Juan National Forest(2),....

Impact of Proposed Pre-training



- Pre-training helps improving the model, especially for the zero-shot setting.

Impact of Pre-training Strategies

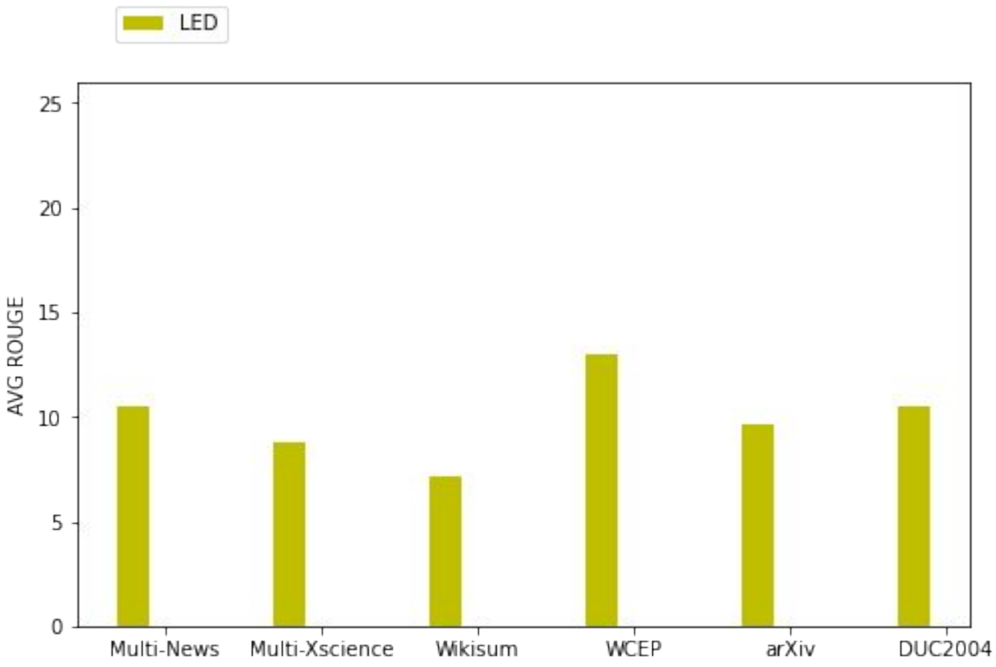


- Same architecture (LED-Base)
- Same input structure
- Same pre-training objective
- Same pre-training dataset
- Zero/Few-shot setting
- **The Entity Pyramid strategy works better than the Principle strategy used in PEGASUS.**

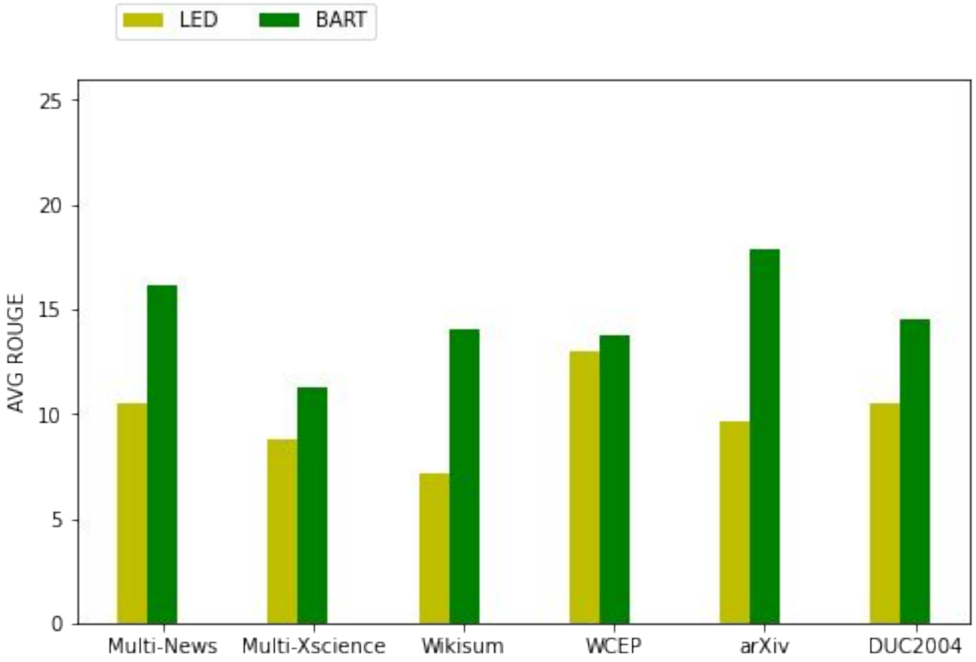
Experiments - Automatic Evaluation

- Evaluation Datasets:
  - Multi-Doc.: Multi-News, Multi-XScience, WCEP, Wikisum, DUC2004
  - Single Doc. arXiv
- Settings:
  - Zero-shot (with length limit)
  - Few-shot: 10/100 training examples, 5 runs for each model
  - Fully supervised
- Compared Models:
  - BART
  - PEGASUS
  - Longformer Encoder Decoder (LED)
  - Prior SOTA Models (fully supervised only)
- Evaluation Metric:
  - ROUGE scores (R-1, R-2, and R-L)

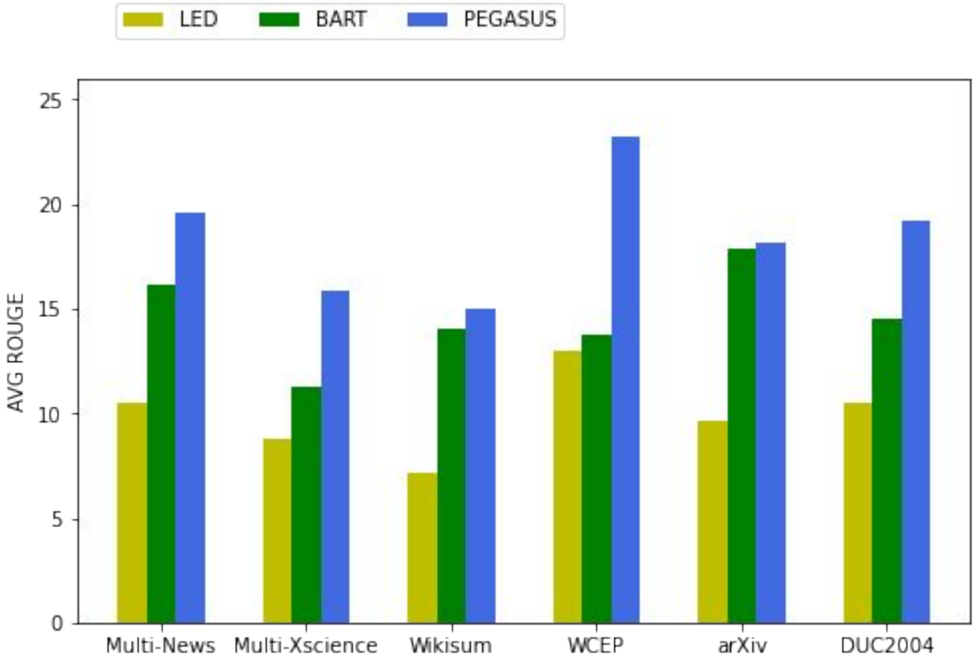
Results on Zero-shot



Results on Zero-shot

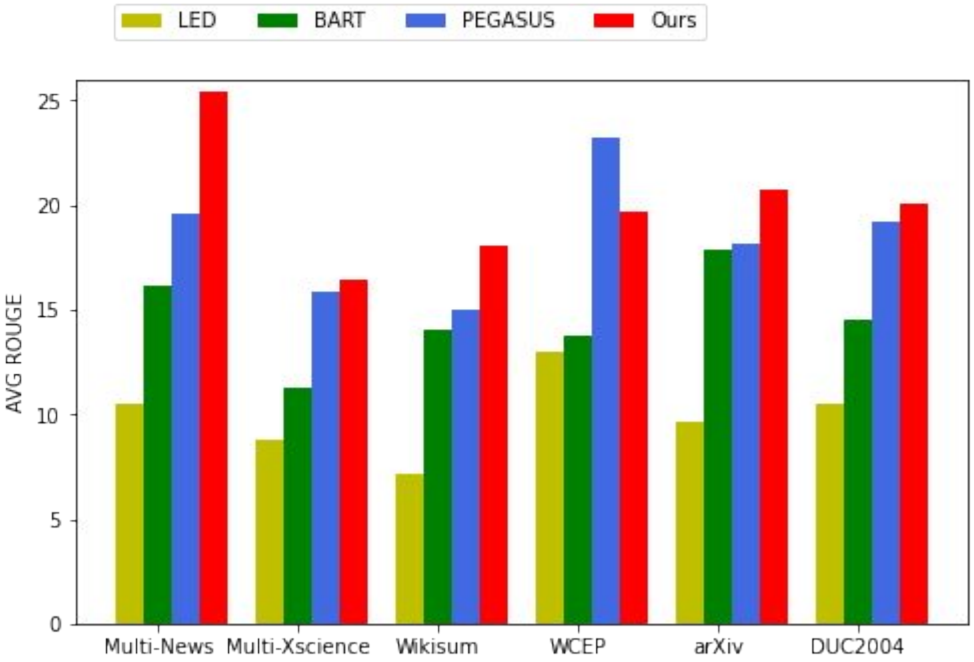


Results on Zero-shot



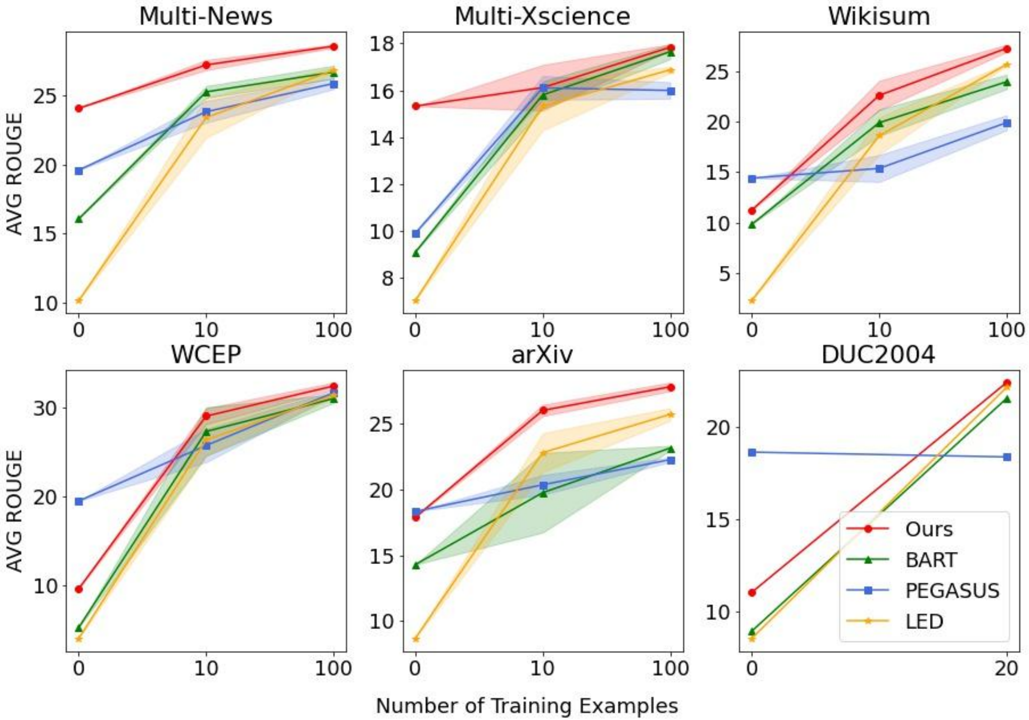
- PEGASUS is also pre-trained for summarization downstream task, thus it performs better than the other two models

Results on Zero-shot



- PEGASUS is also pre-trained for summarization downstream task, thus it performs better than the other two models
- Our model outperforms all the other pre-trained models on most of the datasets (up to 5 ROUGE points)

Results on Few-shot



- Our model outperform all the other pre-trained models on all the datasets.

Results on Fully Supervised

DATASETS	Prev. SOTA			PRIMERA		
	R1	R2	RL	R1	R2	RL
Multi-News	49.2	19.6	24.5	**49.9**	**21.1**	**25.9**
Multi-XScience	**34.1**	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**

- Our model achieves SOTA on several multi-document summarization datasets, as well a single-document summarization dataset.

Experiments - Human Evaluation

- Datasets:
  - DUC 2007
  - TAC 2008
- Metrics:
  - Pyramid Evaluation
  - Fluency (following DUC guidelines*)

* https://www-nlpir.nist.gov/projects/duc/duc2007/quality-questions.txt

Human Evaluation - Pyramid & Fluency

	Model	Pyramid Evaluation				Fluency		
		S_r	R	P	F	Gram.	Ref.	Str.&Coh.
DUC 2007	PEGASUS	6.0	2.5	2.4	2.4	4.45	4.35	1.95
	LED	9.6	3.9	4.0	3.8	4.35	4.50	3.20
	PRIMERA	**12.5**	**5.1**	**5.0**	**5.0**	**4.70**	**4.65**	**3.70**
TAC 2008	PEGASUS	**8.7**	**9.1**	9.4	9.1	**4.40**	4.20	3.20
	LED	6.9	7.1	10.8	8.4	3.10	3.80	2.55
	PRIMERA	8.5	8.9	**10.0**	**9.3**	**4.40**	**4.45**	**4.10**

- PRIMERA also shows a better performance on human evaluation, regarding both pyramid evaluation and fluency evaluation.

Takeaway

- **PRIMERA**, a pre-trained model for **multi-document summarization.**
- It is **pre-trained** with new strategy, **Entity Pyramid.**
- **PRIMERA** reduces the need for dataset-specific architectures and large labeled data.
- **PRIMERA** achieves SOTA on multiple datasets under zero/few-shot and fully supervised, and shows advantage in human evaluation.
- Sample Usage:

```
from transformers import AutoTokenizer, LEDForCondiionalGeneration

Tokenizer = AutoTokenizer.from_pretrained(‘allenai/PRIMERA’)
Model = LEDForConditionalGeneration.from_pretrained(‘allenai/PRIMERA’)
```

Code can be found here: https://github.com/allenai/PRIMER

Future Work

- Controllable generator to better control the length of generated summaries for zero-shot setting
- Evaluate PRIMERA and its Pyramid Entity strategy on other tasks with multiple documents as input, e.g. Multi-hop QA

Thanks!