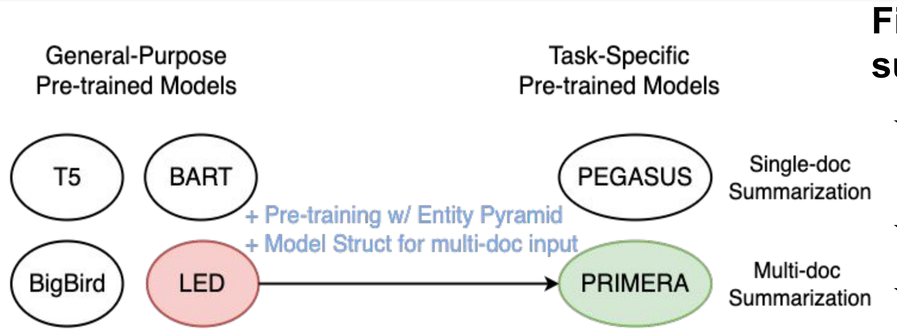
PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization

Wen Xiao[1], **Iz Beltagy**[2], **Giuseppe Carenini**[1], **Arman Cohan**[2]
xiaowen3@cs.ubc.ca

[1]University of British Columbia

[2]Allen Institute of AI

1. Motivation & Highlights



First pre-trained model focused on multi-document summarization

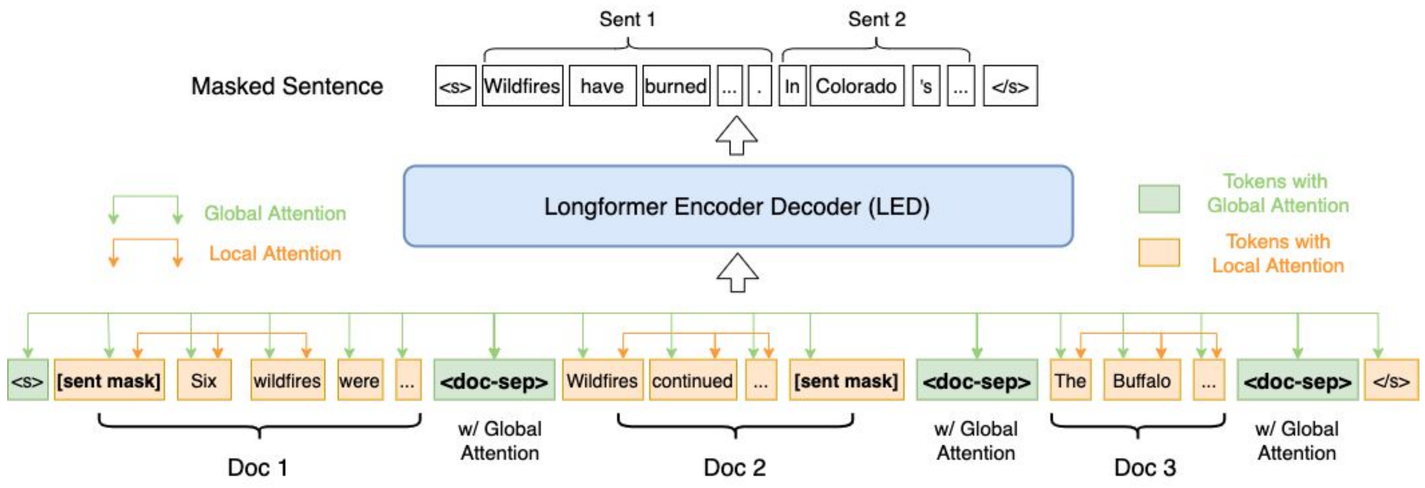
- ★ Generally easier to use than the models with custom architectures.
- ★ Able to work with limited data (as low as 10 examples).
- ★ SOTA results in data rich setting (fully-supervised)

Sample Usage (HF)

```
from transformers import AutoTokenizer,\
    LEDForConditionalGeneration

tokenizer = AutoTokenizer.\
    from_pretrained("allenai/PRIMERA")
model = LEDForConditionalGeneration.\
    from_pretrained("allenai/PRIMERA")
```

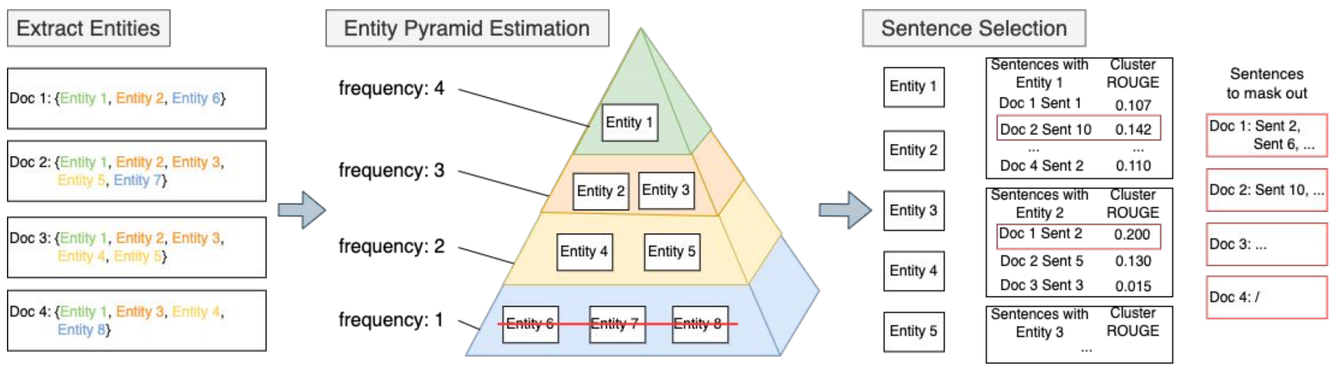
2. Input Structure & Architecture



- **Input Structure:**
  - Simply concatenate documents into one sequence
  - Special document separator tokens (<doc-sep>)
- **Architecture: Longformer Encoder Decoder**
  - Global + Local Attention
  - Allows for long sequence input
  - Global Attention on <doc-sep>

3. Pre-training w/ Entity Pyramid Masking Strategy

- **Goal:** Teach model to connect/aggregate information across a "cluster" of related docs
- **Multi-doc corpus:** Newshead (360k clusters, 3.5 doc/cluster on average)
- **Objective**: **G**ap **S**entence **G**eneration (like in PEGASUS[1])
  - Select SALIENT sentences (as pseudo-summary) and reconstruct them based on other sentences
- **Novel GSS Strategy**: Entity Pyramid (inspired by Pyramid Evaluation [2])
  - Select sentences that best represent the entire cluster of input documents, based on the information overlap with all the other documents in the cluster



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

An example of how the *Entity Pyramid strategy* selects more representative sentences than the Principle strategy

Document #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.

Document #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*...

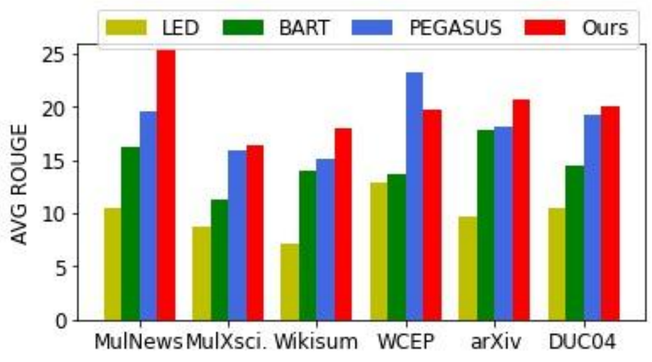
Document #3 The Buffalo Fire west of Denver is ... Several wildfires in Colorado have prompted thousands of home evacuations ...(0.172)... Nearly 1,400 homes have been evacuated in Summit County, Colorado, ...(0.179)... "Under current conditions, one abandoned campfire or spark could cause a catastrophic wildfire, ... , with human life and property," said Richard Bustamante, SJNF forest fire staff officer ...

Entities with High Frequency

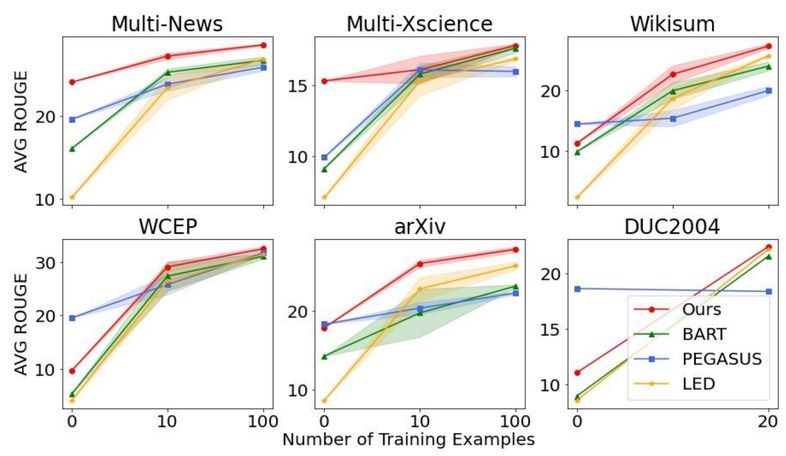
Colorado, 416, Tuesday, Wildfires, San Juan National Forest,...

4. Top performer in zero-/few-shot & fully supervised settings & human evaluation

a. Zero-shot Results



b. Few-shot Results



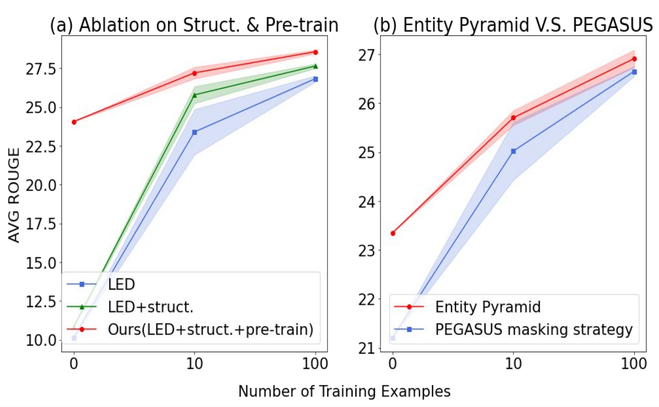
c. Fully Supervised Results

DATASETS	Prev. SOTA			Ours(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**

d. Human Evaluation Results (DUC 2007 & TAC 2008)

Model	Pyramid Evaluation				Fluency		
	S_r	R	P	F	Gram	Ref.	Str.&Coh
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
Ours(PRIMERA)	**12.5**	**5.1**	**5.0**	**5.0**	**4.70**	**4.65**	**3.70**
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
Ours(PRIMERA)	8.5	8.9	**10.0**	**9.3**	**4.40**	**4.45**	**4.10**

e. Ablation Study on Multi-News



a. In zero-shot setting, **PRIMERA** outperforms all the other pre-trained models on most of the datasets (up to 5 ROUGE points)
b. In few-shot setting, **PRIMERA** outperforms all the other pre-trained models on all the datasets.
c. In fully supervised setting, **PRIMERA** achieves SOTA on multi-doc/single doc summarization datasets
d. **PRIMERA** also shows a better performance on human evaluation
e. The ablation study shows both **input structure** and **Entity Pyramid strategy** help improve the performance

[1] Evaluating Content Selection in Summarization: The Pyramid Method. Nenkova & Passonneau, NAACL 2004
[2] PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. Zhang et al. ICML 2020

ACL 2022