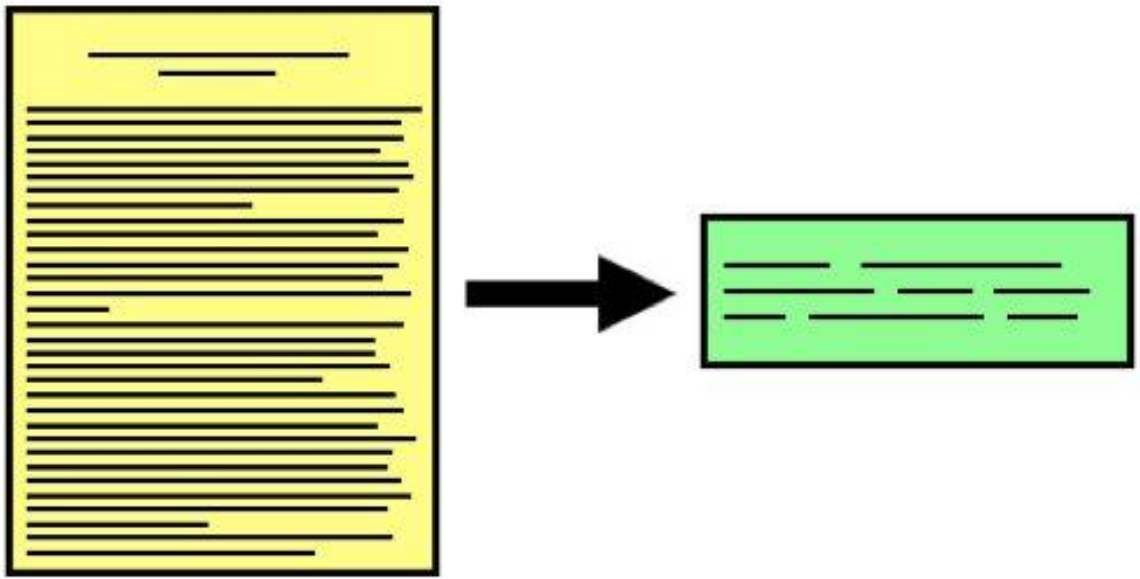
PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization

Guided By : **Prof. Brijesh Bhatt**
Prof. Siddharth Shah

Prepared By : **Darshan Tank**
(8th Sem)

- **What is summarization ?**

- **What are the types of summarization ?**
  - **Abstractive**
  - **Extractive**



- **How machine gonna achieve this task ?**
  - **Seq2Seq**
    - **RNN (LSTM, GRUs).**
    - **Encoder-Decoder mechanism, Context vector.**

- **Issue With RNN based model**
  - **What if we have powerful machine with high class GPUs. Will we able to utilize that power of machine with RNN.**
  - **Loss of information Gain**
  - **Vanishing Gradient**

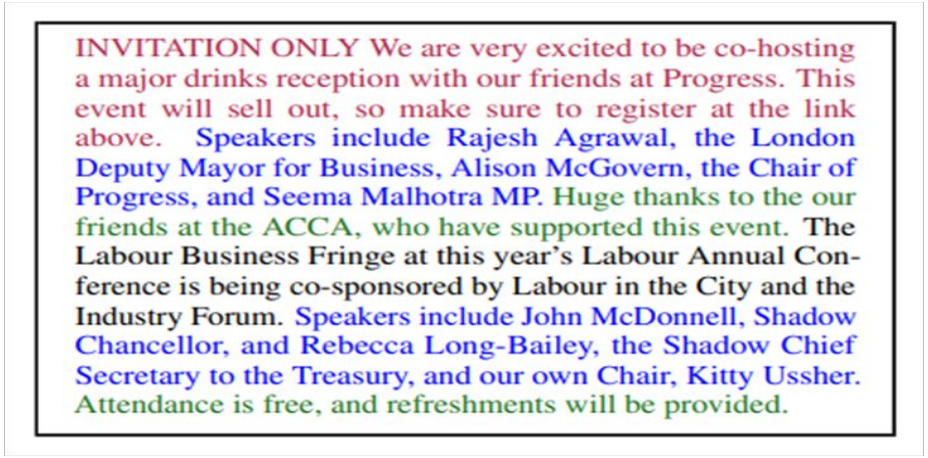
- **Transformation from RNN to TRANSFORMER.**
  - **Parallel Computing.**
  - **Self-Attention with multiple-Head.**
  - **Positional Encoding.**

- **Transformers are Data Hungry !**
  - **Introduction to Pre-training and Fine-tuning.**

- **What are Pre-training models?**
  - **MASS**
  - **UniLM**
  - **T5**
  - **BART**
  - **MLM**

- **Will these are enough to Perform Abstractive Text Summarization?**
  - **Introduction to GSG(Gap-Sentence-Generation)**
    - **How to select Gap-Sentences?**
      - **Randomly**
      - **Lead**
      - **Principal**
        - **Independently**
        - **Sequentially**

TRANSFORMER

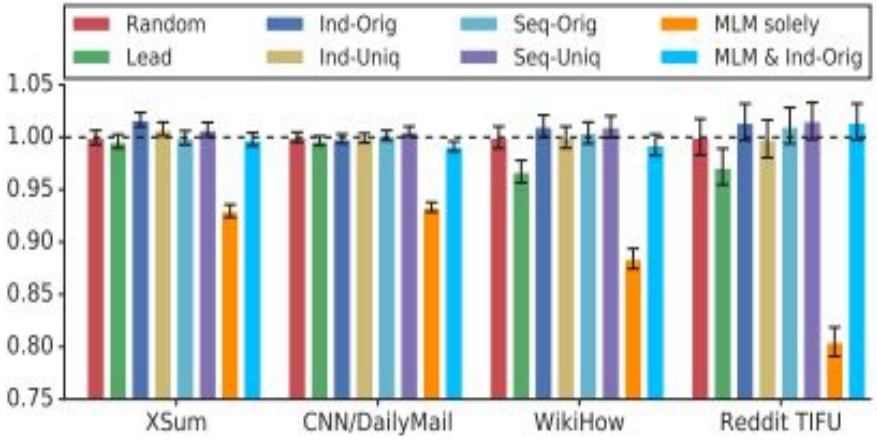

INVITATION ONLY We are very excited to be co-hosting a major drinks reception with our friends at Progress. This event will sell out, so make sure to register at the link above. Speakers include Rajesh Agrawal, the London Deputy Mayor for Business, Alison McGovern, the Chair of Progress, and Seema Malhotra MP. Huge thanks to the our friends at the ACCA, who have supported this event. The Labour Business Fringe at this year's Labour Annual Conference is being co-sponsored by Labour in the City and the Industry Forum. Speakers include John McDonnell, Shadow Chancellor, and Rebecca Long-Bailey, the Shadow Chief Secretary to the Treasury, and our own Chair, Kitty Ussher. Attendance is free, and refreshments will be provided.

Random

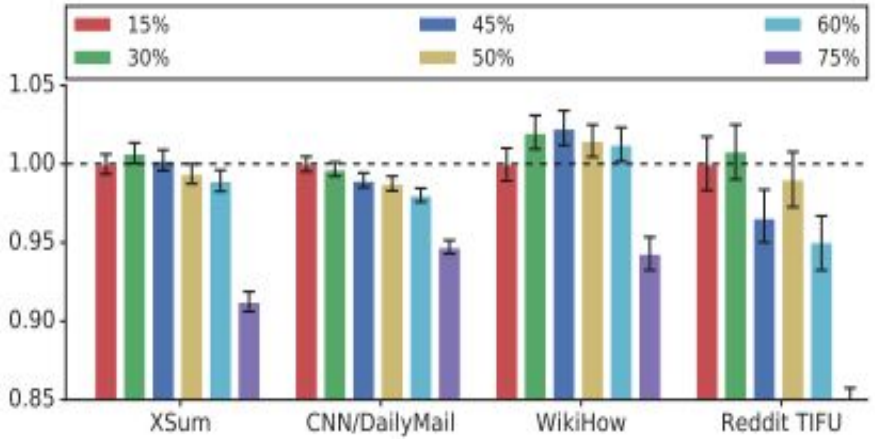
Lead

Principal

- **How many sentences should be masked?**
  - **GSR(Gap Sentences Ration)**



(a) Effect of pre-training objectives (30% GSR).



(b) Effect of gap sentences ratio with GSG (Ind-Orig).

From the diagrams we can see that **GSR value 15% & 30%** works well for all the given data set.

Now, We have everything with us let's start experiment.

	Batch Size	Parameters	Data-Set	Number of layers for encoder-decoder(L)	Hidden layer size(H)	Feed-Forward layer size(F)	NUmber of Self-Attention head(A)
$PEGASUS_{BASE}$	256	223M	4	12	768	3072	12
$PEGASUS_{LARGE}$	8192	568M	12	16	1024	4096	16

Vocabulary Option = Unigram with size of 96k

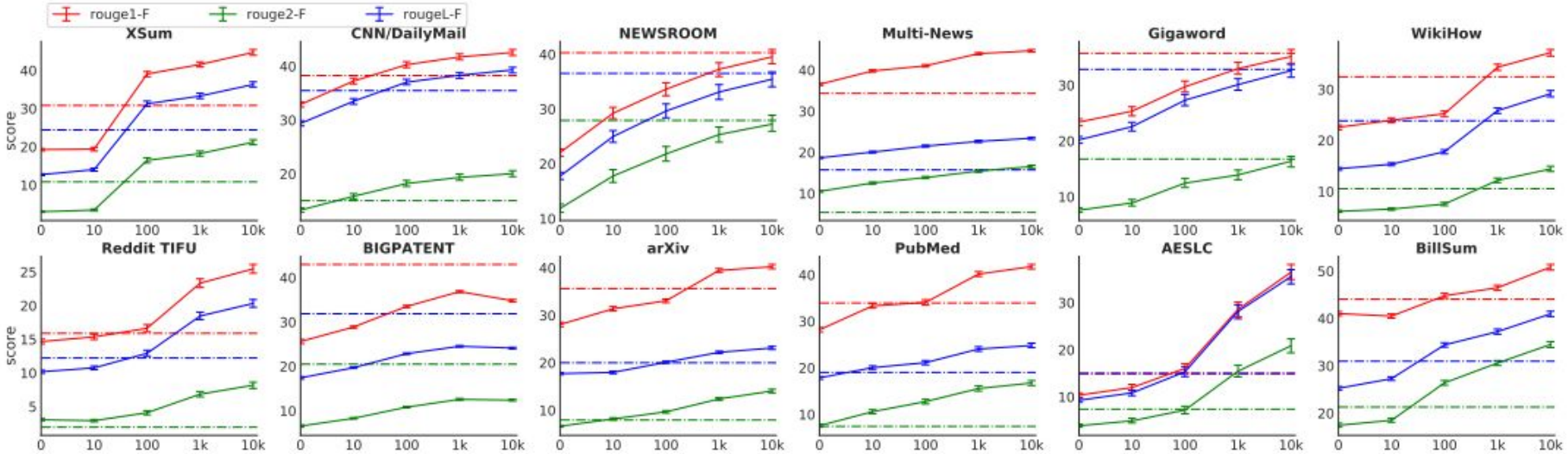
Gradient descent = For optimization, both pre-training and fine-tuning used **Adafactor** (Shazeer & Stern, 2018) with **square root learning rate decay** and **dropout rate of 0.1.**

Final Output :

R1/R2/RL	Dataset size	$\text{Transformer}_{\text{BASE}}$	$\text{PEGASUS}_{\text{BASE}}$	Previous SOTA	$\text{PEGASUS}_{\text{LARGE}}$ (C4)	$\text{PEGASUS}_{\text{LARGE}}$ (HugeNews)
XSum	226k	30.83/10.83/24.41	39.79/16.58/31.70	45.14/22.27/37.25	45.20/22.06/36.99	**47.21/24.56/39.25**
CNN/DailyMail	311k	38.27/15.03/35.48	41.79/18.81/38.93	**44.16**/21.28/40.90	43.90/21.20/40.76	**44.17/21.47/41.11**
NEWSROOM	1212k	40.28/27.93/36.52	42.38/30.06/38.52	39.91/28.38/36.87	**45.07/33.39/41.28**	**45.15/33.51/41.33**
Multi-News	56k	34.36/5.42/15.75	42.24/13.27/21.44	43.47/14.89/17.41	46.74/17.95/24.26	**47.52/18.72/24.91**
Gigaword	3995k	35.70/16.75/32.83	36.91/17.66/34.08	**39.14/19.92/36.57**	38.75/**19.96**/36.14	**39.12/19.86**/36.24
WikiHow	168k	32.48/10.53/23.86	36.58/15.64/30.01	28.53/9.23/26.54	**43.06/19.71/34.80**	41.35/18.51/33.42
Reddit TIFU	42k	15.89/1.94/12.22	24.36/6.09/18.75	19.0/3.7/15.1	**26.54/8.94/21.64**	**26.63/9.01/21.60**
BIGPATENT	1341k	42.98/20.51/31.87	43.55/20.43/31.80	37.52/10.63/22.79	**53.63/33.16/42.25**	53.41/32.89/42.07
arXiv	215k	35.63/7.95/20.00	34.81/10.16/22.50	41.59/14.26/23.55	**44.70/17.27/25.80**	**44.67/17.18/25.73**
PubMed	133k	33.94/7.43/19.02	39.98/15.15/25.23	40.59/15.59/23.59	**45.49/19.90/27.69**	45.09/19.56/27.42
AESLC	18k	15.04/7.39/14.93	34.85/18.94/34.10	23.67/10.29/23.44	**37.69/21.85/36.84**	37.40/21.22/36.45
BillSum	24k	44.05/21.30/30.98	51.42/29.68/37.78	40.80/23.83/33.73	**57.20**/39.56/**45.80**	**57.31/40.19/45.82**

R1/R2/RL	XSum	CNN/DailyMail	Gigaword
BERTShare (Rothe et al., 2019)	38.52/16.12/31.13	39.25/18.09/36.45	38.13/19.81/35.62
MASS (Song et al., 2019)	39.75/17.24/31.95	42.12/19.50/39.01	38.73/19.71/35.96
UniLM (Dong et al., 2019)	-	43.33/20.21/40.51	38.45/19.45/35.75
BART (Lewis et al., 2019)	45.14/22.27/37.25	**44.16**/21.28/40.90	-
T5 (Raffel et al., 2019)	-	43.52/**21.55**/40.69	-
$\text{PEGASUS}_{\text{LARGE}}$ (C4)	45.20/22.06/36.99	43.90/21.20/40.76	38.75/**19.96/36.14**
$\text{PEGASUS}_{\text{LARGE}}$ (HugeNews)	**47.21/24.56/39.25**	**44.17/21.47/41.11**	**39.12/19.86/36.24**

- **Difficulty to collect a large number of supervised examples to train or fine-tune a summarization model.**
  - **low-resource summarization setting** (picked the first 10^k (k = 1, 2, 3, 4) training examples from each dataset to fine-tune $\text{PEGASUS}_{\text{LARGE}}$).



Solid lines → $\text{PEGASUS}_{\text{LARGE}}$

Dashed lines → $\text{Transformer}_{\text{BASE}}$ models (equivalent in capacity as $\text{PEGASUS}_{\text{BASE}}$)

Conclusion :

Goal : Abstractive Summarization

Solution : PEGASUS

Pre-training Data-set	: C4 and HugeNews
Fine-tuning Data-set	: List of 12 Data-sets
DownStream Task	: Abstractive Summarization
Pre-training Model	: GSG(Gap-Sentence-Generation)
Which Gap-sentences	: Principal approach (Independent-Original)
GSR value	: 15%
Gradient Descent Algo	: Adafactor
Vocabulary Size	: Unigram with size of 96k