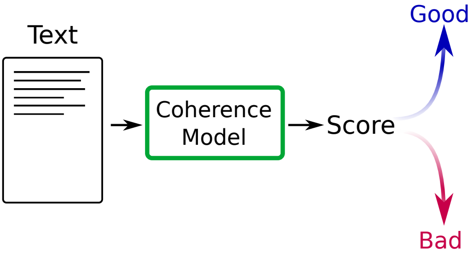
Rethinking Self-Supervision Objectives for Generalizable Coherence Modeling

Prathyusha Jwalapuram[†] **Shafiq Joty**[†§] **Xiang Lin**[†]
[†]Nanyang Technological University, Singapore
[§]Salesforce Research Asia, Singapore

Coherence Modeling



- Increasing claims of fluency - applications in language generation, summarization, machine translation, etc.
- Most work on coherence modeling **ignores downstream applications**

Outline

1. Introduction
2. Methodology
   - Contrastive Training
   - Hard Negative Mining
   - Global Negative Queue
3. Experiments
4. Analysis
5. Conclusions

Motivation

Original Document	
(S1)	Judy and I were in our back yard when the lawn started rolling like ocean waves.
(S2)	We ran into the house to get Mame, but the next tremor threw me in the air and bounced me as I tried to get to my feet.
(S3)	We are all fine here, although Mame was extremely freaked.
(S4)	Books and tapes all over my room.
(S5)	Not one thing in the house is where it is supposed to be, but the structure is fine.
Permuted Document	
(S4)	Books and tapes all over my room.
(S3)	We are all fine here, although Mame was extremely freaked.
(S2)	We ran into the house to get Mame, but the next tremor threw me in the air and bounced me as I tried to get to my feet.
(S5)	Not one thing in the house is where it is supposed to be, but the structure is fine.
(S1)	Judy and I were in our back yard when the lawn started rolling like ocean waves.

- Coherence models are commonly trained and evaluated on the **permuted document task** [Barzilay and Lapata, 2005]

Motivation

- Performance on permuted document task only partially indicative of coherence modeling capabilities [Pishdad et al., 2020]
- SOTA models perform well on permuted document task but generalize poorly to downstream tasks [Mohiuddin et al., 2021]

Method

- Coherence models usually trained **pairwise** on permuted document task
  - Model only exposed to limited number of samples in this setting [Li and Jurafsky, 2017]

Method

- Coherence models usually trained **pairwise** on permuted document task
  - Model only exposed to limited number of samples in this setting [Li and Jurafsky, 2017]
- Learning with more negatives maximizes the mutual information between representations [van den Oord et al., 2018]

Method

- Coherence models usually trained **pairwise** on permuted document task
  - Model only exposed to limited number of samples in this setting [Li and Jurafsky, 2017]

- Learning with more negatives maximizes the mutual information between representations [van den Oord et al., 2018]

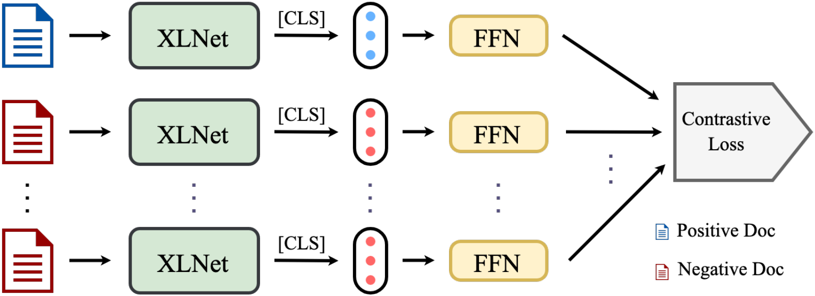
⇒ Compare each 'positive' document to multiple 'negative' documents using **contrastive learning** [Gutmann and Hyvärinen, 2010]

Model

- Obtain $[\mathrm{CLS}]$ representation of input document $\mathcal{D}$ using XLNet [Yang et al., 2019]
- Linear layer converts document representation to coherence score $f_\theta(\mathcal{D})$

$\rightarrow$ No task-specific architecture - trained purely through **self-supervision**

Contrastive Learning



Hard Negative Mining

- Quality of negatives used in contrastive training strongly influences model success [Wu et al., 2020]

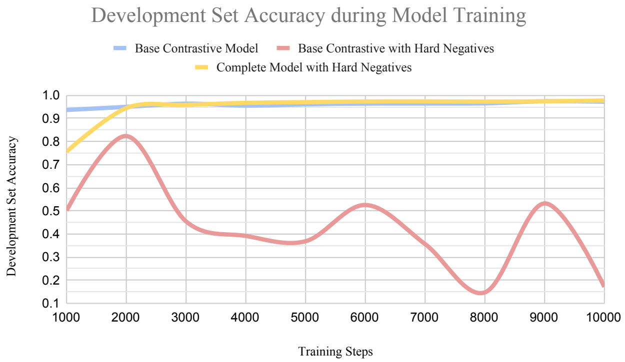
  $\Rightarrow$ Perform hard negative mining

Local Negative Sample Ranking

- Sample more negatives than needed for training ($h > N$)
- Train model for a few steps
- Score the h negatives for the next set of training data
- Use top N to train the next steps

$\rightarrow$ Model iteratively mines harder and harder samples as it improves

Hard Negative Mining



- Training with hardest negatives can lead to bad local minima [Xuan et al., 2020]
- Larger gradient norms result in abrupt gradient steps [Xiong et al., 2020]

Outline

Global Negative Queue

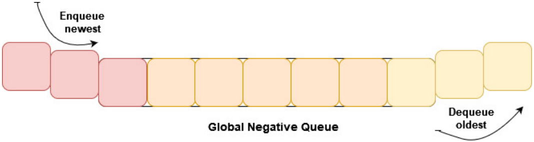
- Number of negatives for contrastive training limited by resource constraints

  $\rightarrow$ Maintain large global queue of negative samples independent of current training sample

Global Negative Queue

- Number of negatives for contrastive training limited by resource constraints

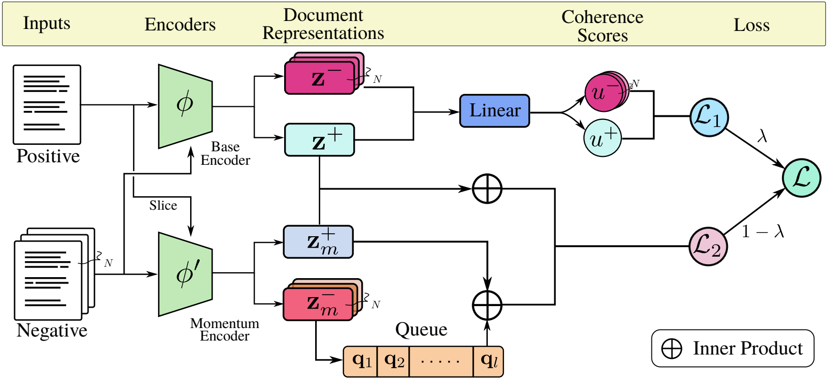
  → Maintain large global queue of negative samples independent of current training sample



- But representations in the queue will become inconsistent as training progresses

  → Use an auxiliary momentum encoder [He et al., 2020]

Model Architecture



Momentum Encoder

- Auxiliary momentum encoder parameters are **not updated** through backpropagation
- Momentum encoder ϕ' is updated based on the base encoder ϕ:

$$\phi' \leftarrow \mu * \phi' + (1 - \mu) * \phi \tag{1}$$

- $\mu \in [0, 1)$ is the **momentum coefficient**

Momentum Encoder

- Use hard negative mining in combination with momentum encoder
- Momentum model - **temporal ensemble** of exponential-moving-average versions of base model
- Due to this, gradients from the momentum loss also help in stabilising the overall training

Test sets

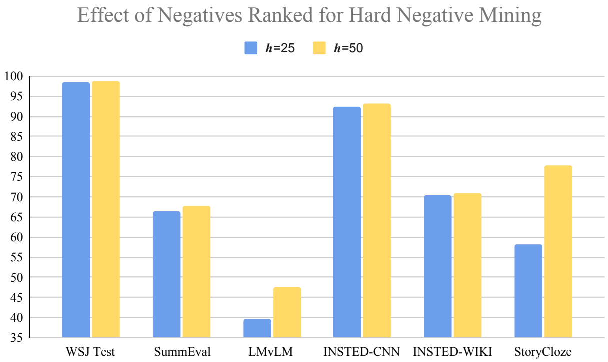
WSJ:	Standard permuted document train & test set
SummEval:	Machine generated summaries [Fabbri et al., 2020]
LMvLM:	Language model output
INSteD-CNN:	Sentence instrusion detection (CNN) [Shen et al., 2021]
INSteD-Wiki:	Sentence intrusion detection (Wikipedia) [Shen et al., 2021]
StoryCloze:	Commonsense reasoning [Sharma et al., 2018]

Results

Model	WSJ	SumEval	LMvLM	INS-CNN	INS-Wiki	StryClz
LCD-G	90.39	54.15	0.419	61.24	55.09	51.76
LCD-I	91.56	51.71	0.420	60.23	53.50	52.69
LCD-L	90.24	53.56	0.404	55.07	51.04	50.09
UNC	94.11	46.28	0.463	67.21	55.97	49.39
Our - Pairwise (No FT)	71.70	54.93	0.421	59.96	53.45	51.69
Our - Pairwise	98.23	64.83	0.458	91.96	70.85	71.84
Our - Contrastive	98.59	66.93	0.468	92.84	71.86	72.83
Our - Full Model	98.58	67.19	0.473	93.36	72.04	74.62

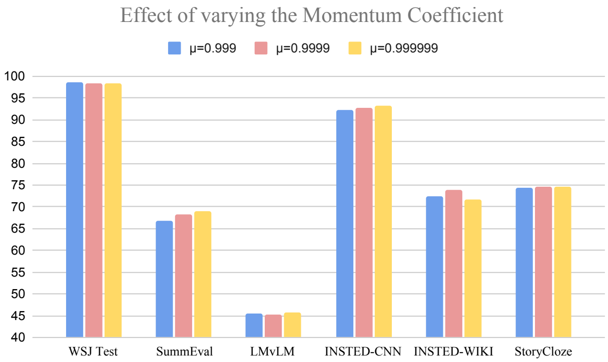
- LCD [Xu et al., 2019] and UNC [Moon et al., 2019] perform poorly across independent test sets
- Our models improve not only on the WSJ test set, but significantly across all the independent test sets

Number of Ranked Negatives



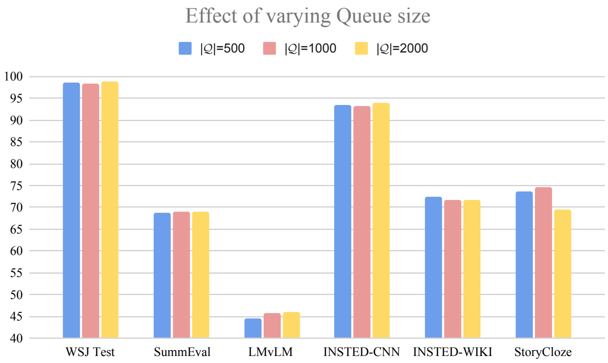
- Increasing number of negatives improves results, particularly on OOD test sets

Momentum Coefficient



- Increasing μ leads to better generalization across independent test sets

Queue Size



- Very high queue size affects generalizability

Varying Task and Dataset

Train Dataset	Neg. Type	Model	WSJ	SummEval	LMvLM	INSteD-CNN	INSteD-Wiki	StoryCloze
INSteD-Wiki	Intrusion	Pairwise	$95.24_{\pm 0.37}$	$53.03_{\pm 1.49}$	$0.490_{\pm 0.01}$	$94.07_{\pm 0.29}$	$82.01_{\pm 0.24}$	$64.21_{\pm 1.98}$
INSteD-CNN	Intrusion	Pairwise	$95.48_{\pm 0.47}$	$57.85_{\pm 2.47}$	$0.502_{\pm 0.01}$	$97.83_{\pm 0.15}$	$73.52_{\pm 1.17}$	$71.75_{\pm 1.81}$
INSteD-Wiki	Permuted	Pairwise	$96.89_{\pm 0.23}$	$64.53_{\pm 0.82}$	$0.491_{\pm 0.01}$	$84.17_{\pm 1.50}$	$71.35_{\pm 0.88}$	$69.09_{\pm 2.29}$
INSteD-CNN	Permuted	Pairwise	$97.03_{\pm 0.12}$	$66.63_{\pm 0.97}$	$0.483_{\pm 0.01}$	$92.61_{\pm 0.62}$	$69.88_{\pm 0.64}$	$68.95_{\pm 1.02}$
WSJ	Permuted	Pairwise	$98.23_{\pm 0.20}$	$64.83_{\pm 1.03}$	$0.458_{\pm 0.02}$	$91.96_{\pm 1.09}$	$70.85_{\pm 1.85}$	$71.84_{\pm 2.33}$

- Overall, training on **WSJ** permuted document task generalizes better than other tasks and datasets

Conclusions

- Increasing ratio and quality of negative samples improves generalizability of the coherence model
- New standard for coherence model evaluation - test the model on several downstream applications
- Encourage research in this new paradigm of coherence modeling

Scan QR code for full paper and code

Thank you!