Self-supervision with more negative samples is better than task-specific architecture for **coherence modeling.**

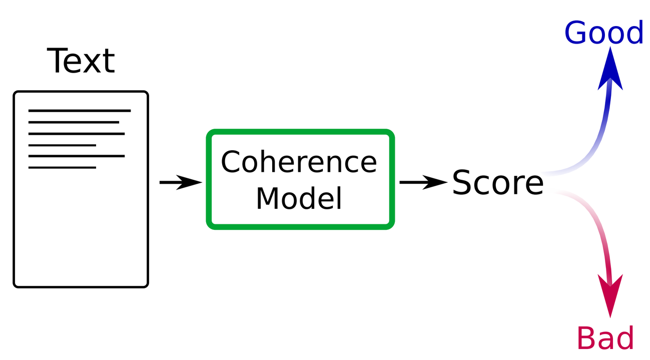
Rethinking Self-Supervision Objectives for Generalizable Coherence Modeling

Prathyusha Jwalapuram†, Shafiq Joty†§ and Xiang Lin†

†Nanyang Technological University, Singapore

§Salesforce Research Asia, Singapore

Introduction



- Increasing claims of fluency - applications in language generation, summarization, MT, etc.
- Most work **ignores downstream applications**
- Typically trained **pairwise** on the permuted document task

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.

- Only partially indicative of coherence modeling [Pishdad et al., 2020]
- SOTA generalizes poorly to downstream tasks [Mohiuddin et al., 2021]

Contrastive Learning

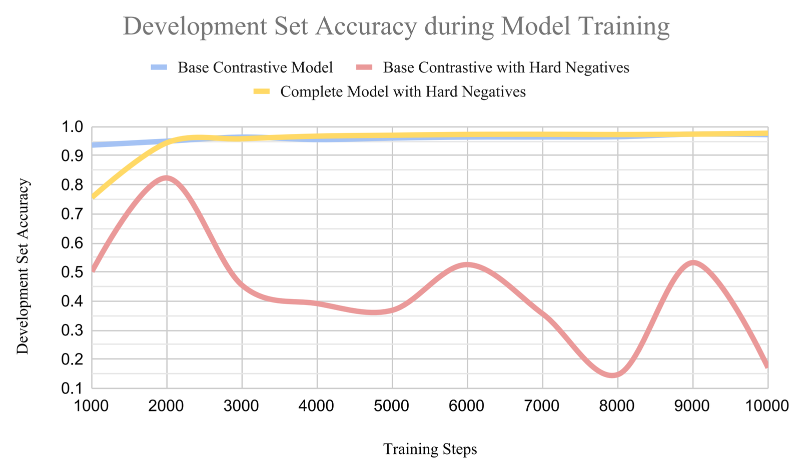
- Maximize mutual information by using **contrastive learning**

  $\Rightarrow$ Compare positive document to **multiple** negative documents

Hard Negative Mining

- Mine **hard negatives** locally
- Sample more than needed and score training samples ahead
- Take top N to train the next steps

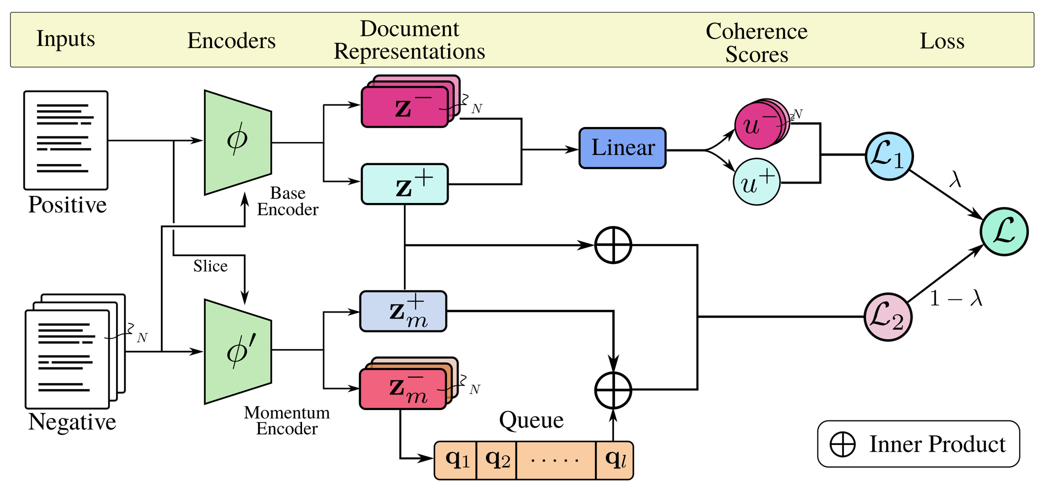
  $\rightarrow$ Causes instablity in training



Auxiliary Momentum Encoder

- Number of negative samples in contrastive training limited by resource constraints

  $\Rightarrow$ Maintain large independent **global queue** of negative samples
- Encode using auxiliary **momentum encoder** to keep representations consistent (not backpropagated through)



- **Temporal ensemble** of exponential-moving-average versions of the base encoder
- Stabilizes hard negative training

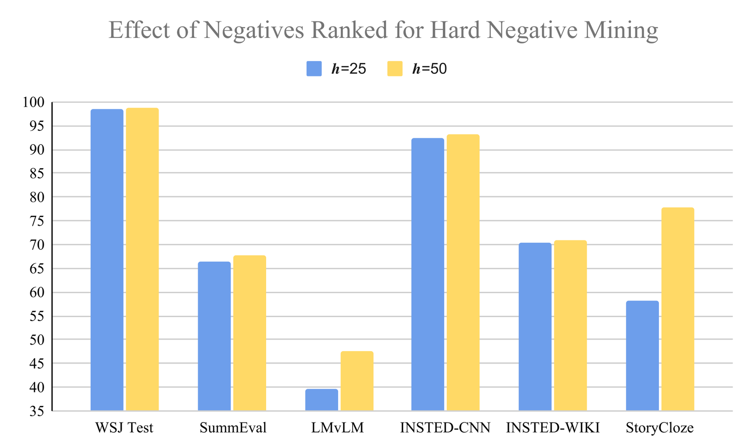
Test Sets

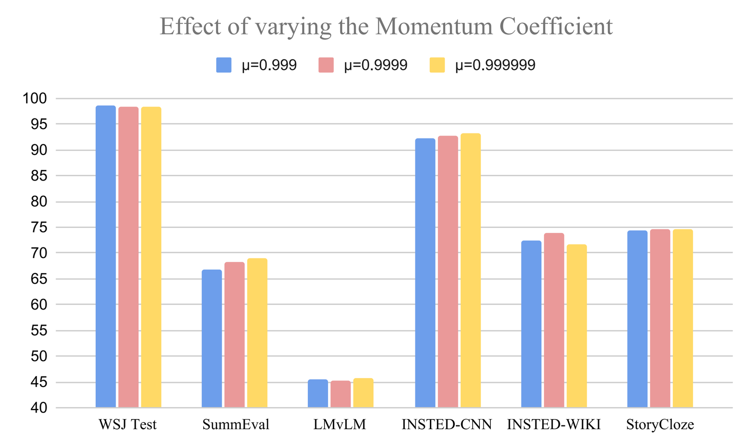
WSJ:	Standard permuted document train & test set
SummEval:	Machine generated summaries [Fabbri et al., 2020]
LMvLM:	Language model output
INSteD-CNN:	Sentence instrusion detection (CNN) [Shen et al., 2021]
INSteD-Wiki:	Sentence intrusion detection (Wikipedia) [Shen et al., 2021]
StoryCloze:	Commonsense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

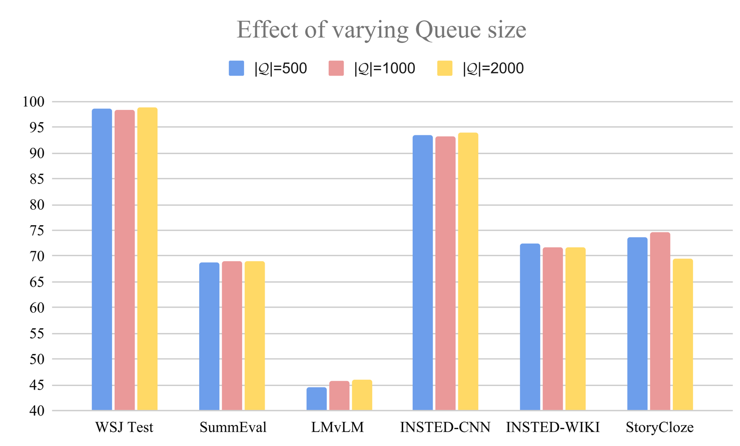
Results

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

Analysis







Conclusions

- Increasing ratio and quality of negative samples improves generalizability
- New standard for coherence model evaluation
- Encourage research in new paradigm of coherence modeling

Scan for full paper and code

NTU NLP