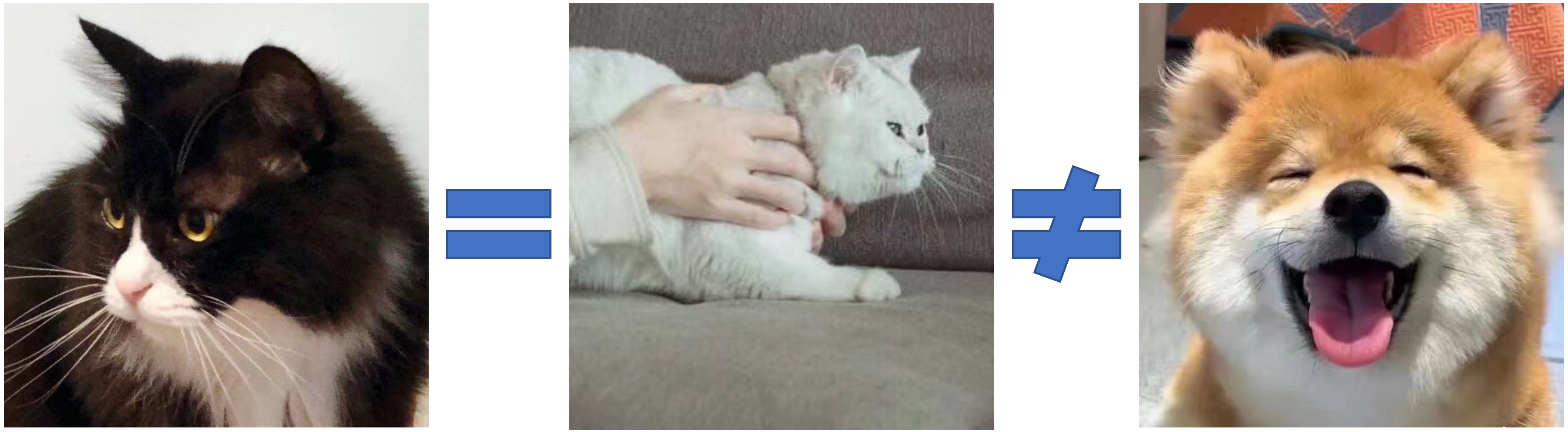
Contrastive Learning on Abstractive Summarization

Qi Jia

Contents

Contrastive Learning

Contrastive learning is a machine learning technique used to learn the **general features** of a dataset **without labels** by teaching the model *which pairs of data points are similar or different*.



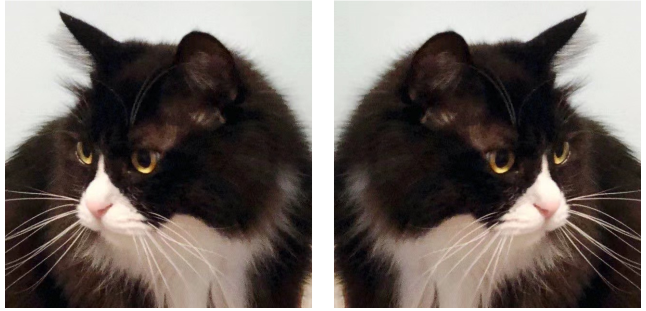
Basic Steps for Contrastive Learning

Step-1: **Define** positive pairs [and negative pairs]

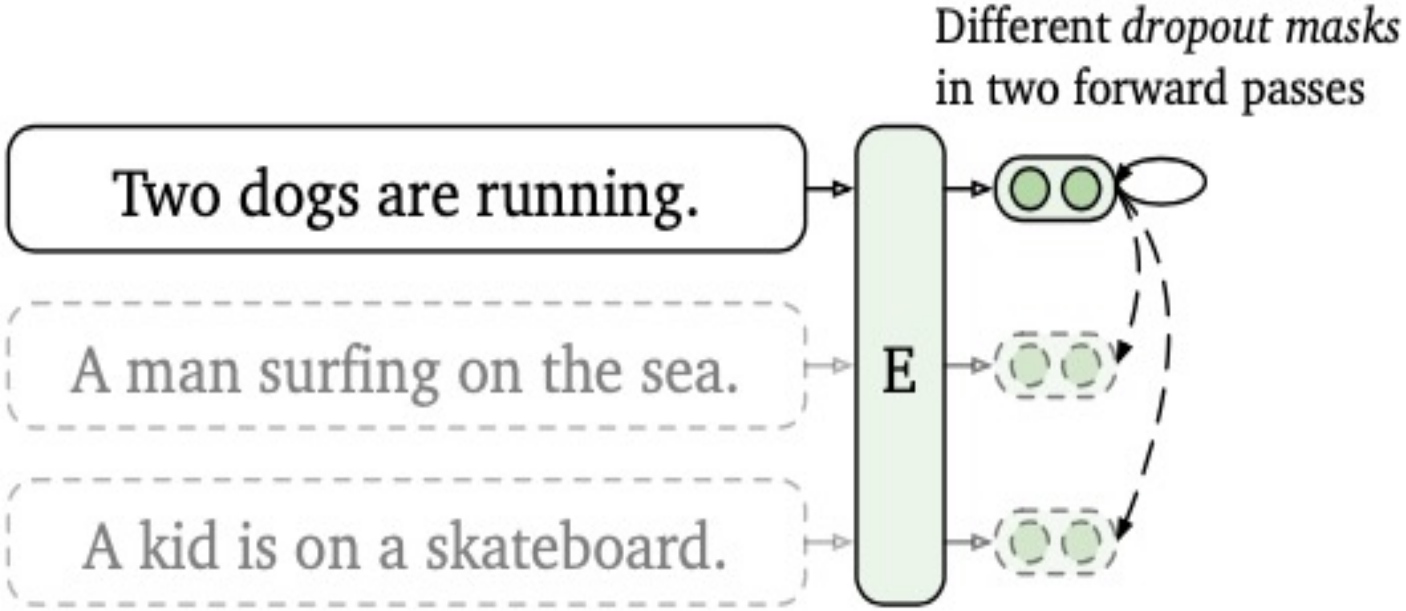
CV

NLP

POS:



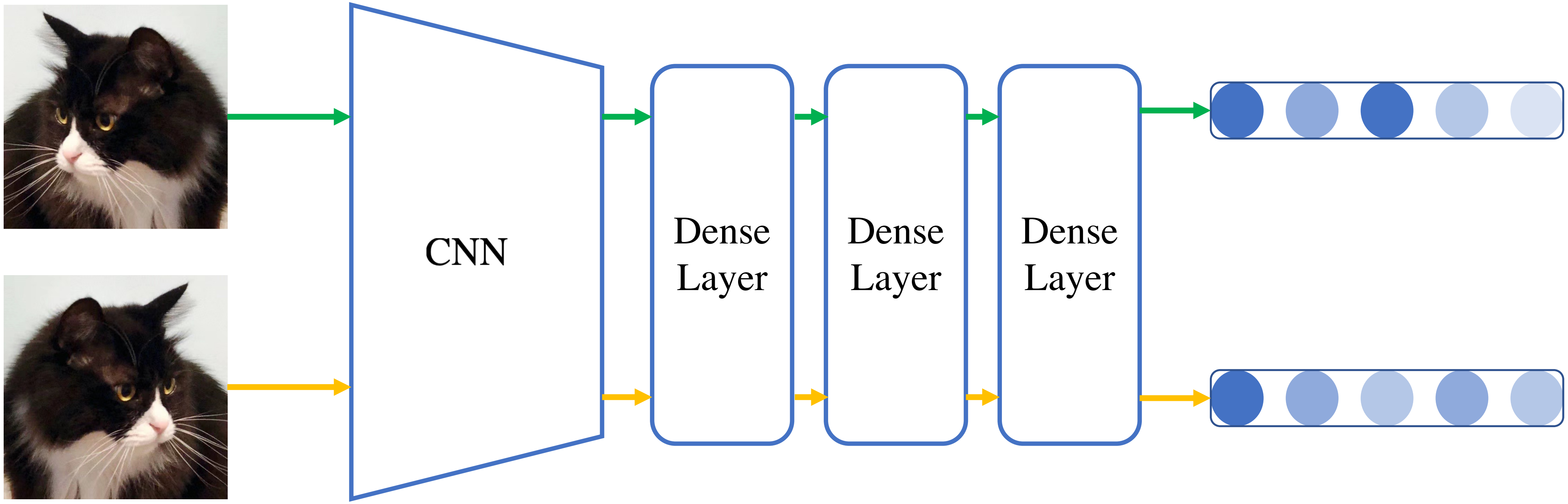
NEG:



Tianyu Gao, Xingcheng Yao, Danqi Chen: SimCSE: Simple Contrastive Learning of Sentence Embeddings. CoRR abs/2104.08821 (2021)

Basic Steps for Contrastive Learning

Step-2: Extract features and get embedded **representations** for pairs by a model.



Basic Steps for Contrastive Learning

Step-3: Train the model to maximize the **similarity** of representations for similar inputs, and minimize the similarity of representations for dissimilar inputs with **a contrastive loss**.

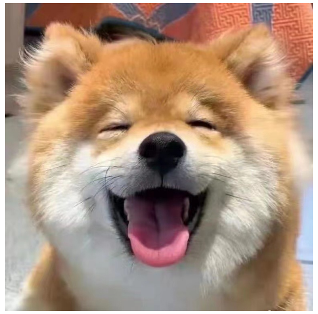
- Distance/Similarity ( , )

- Contrastive Loss e.g. ranking loss

Pairwise/Triplet Ranking Loss

Anchor Sample r_a

Positive Sample r_p



Negative Sample r_n

$$L = \begin{cases} d(r_a, r_p) & positive\ pair \\ \max(0, m - d(r_a, r_n)) & negative\ pair \end{cases}$$

$$L = \max(0, m + d\left(r_a, r_p\right) - d(r_a, r_n))$$

- m: a margin

Other Names used for Ranking Losses

- **Ranking Loss**
  - The *information retrieval* field, where we want to train models to rank items in a specific order.
- **Margin Loss**
  - Use a *margin* to compare the distances between sample representations.
- **Contrastive Loss**
  - The losses are computed contrasting *two* data point representations.
- **Hinge Loss**
  - A similar formulation in the sense it optimizes until a *margin*.
- **Triplet Loss**
  - When triplet training pairs are employed

Take-aways

- Contrastive learning is a **self-supervised**, **task-independent** deep learning technique.

- The model learns general features by comparing **pairs of** data points.

- It can be used as an auxiliary task when **labelled** data is **scarce**.

Abstractive Summarization

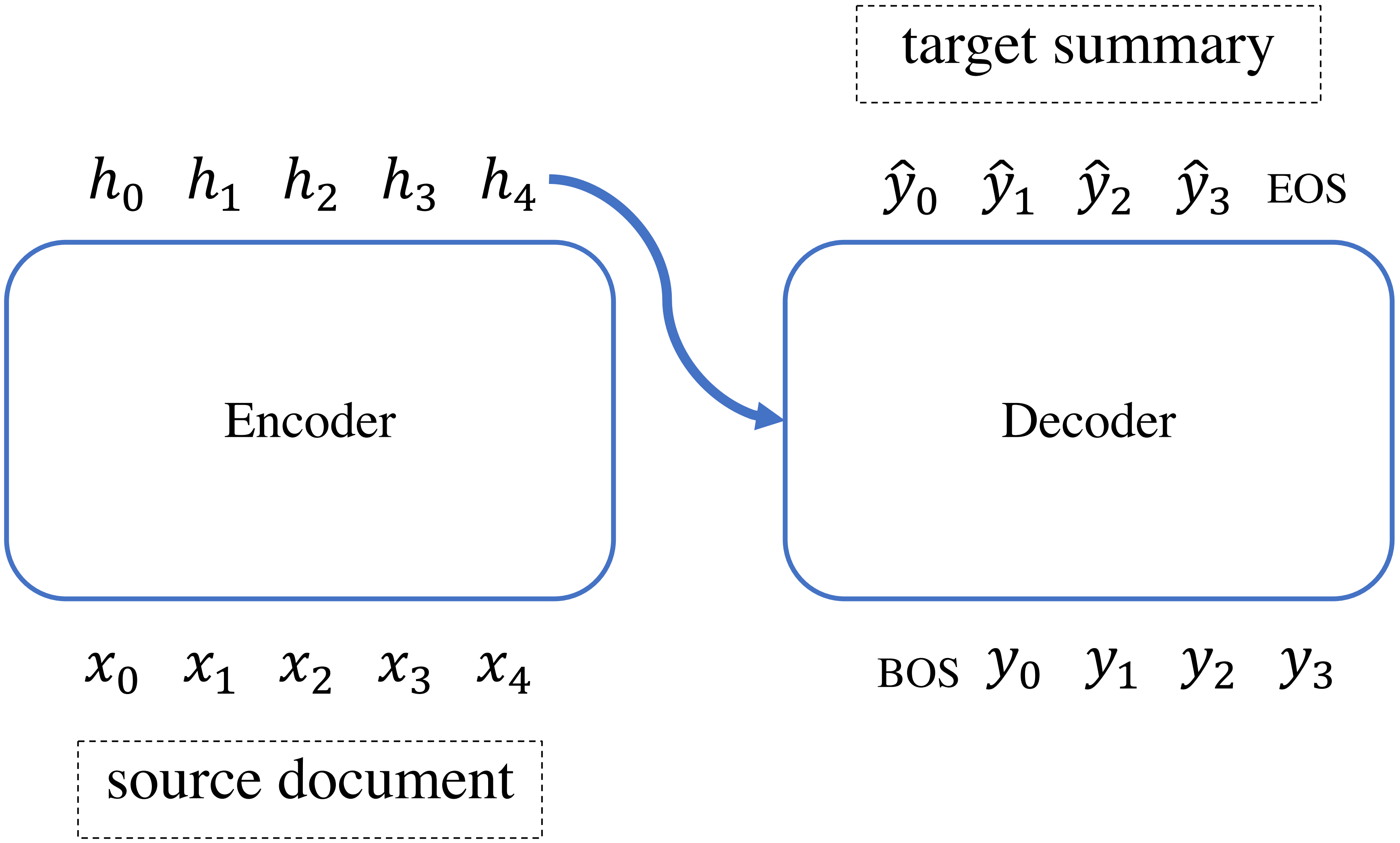
- Source document

justin timberlake and jessica biel , welcome to parenthood . the celebrity couple announced the arrival of their son , silas randall timberlake , in statements to people .`` silas was the middle name of timberlake 's maternal grandfather bill bomar , who died in 2012 , while randall is the musician 's own middle name , as well as his father 's first , '' people reports . the couple announced the pregnancy in january , …

- Abstractive Summary

timberlake and biel welcome son silas randall timberlake . the couple announced the pregnancy in january .

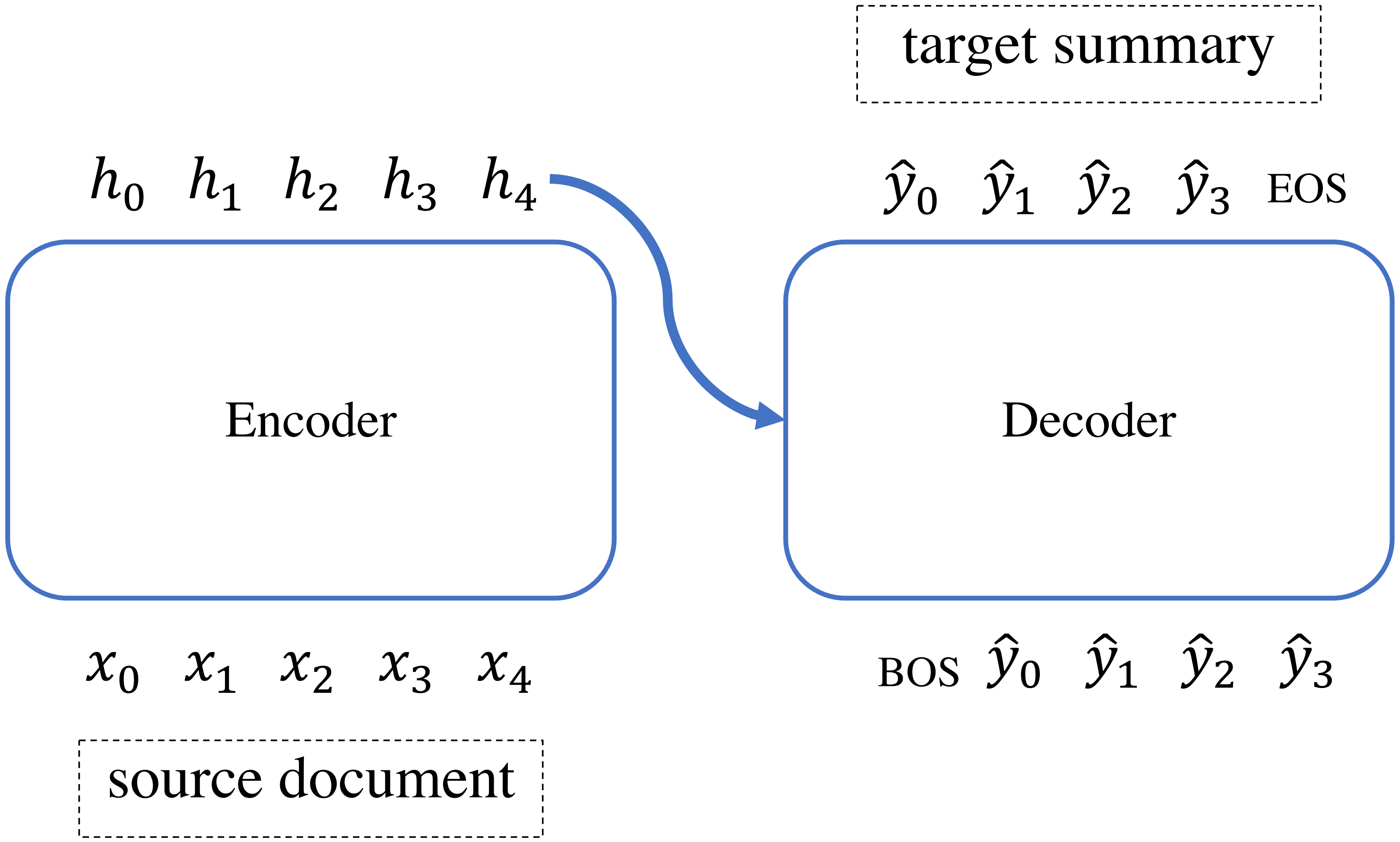
Abstractive Summarization Model



The Negative Log-likelihood Loss

Training

Abstractive Summarization Model



Beam Search

Teacher Forcing

Test

Take-aways

- **Abstractive summarization** is the task of creating a **short**, accurate, and informative **summary** from a **long** text **document** without using the exact sentences from the source.

- Abstractive Summarization Model: Seq-to-Seq **Encoder-decoder** Model Architecture

- Pretrained Language Model: **BART**、**PEGASUS**

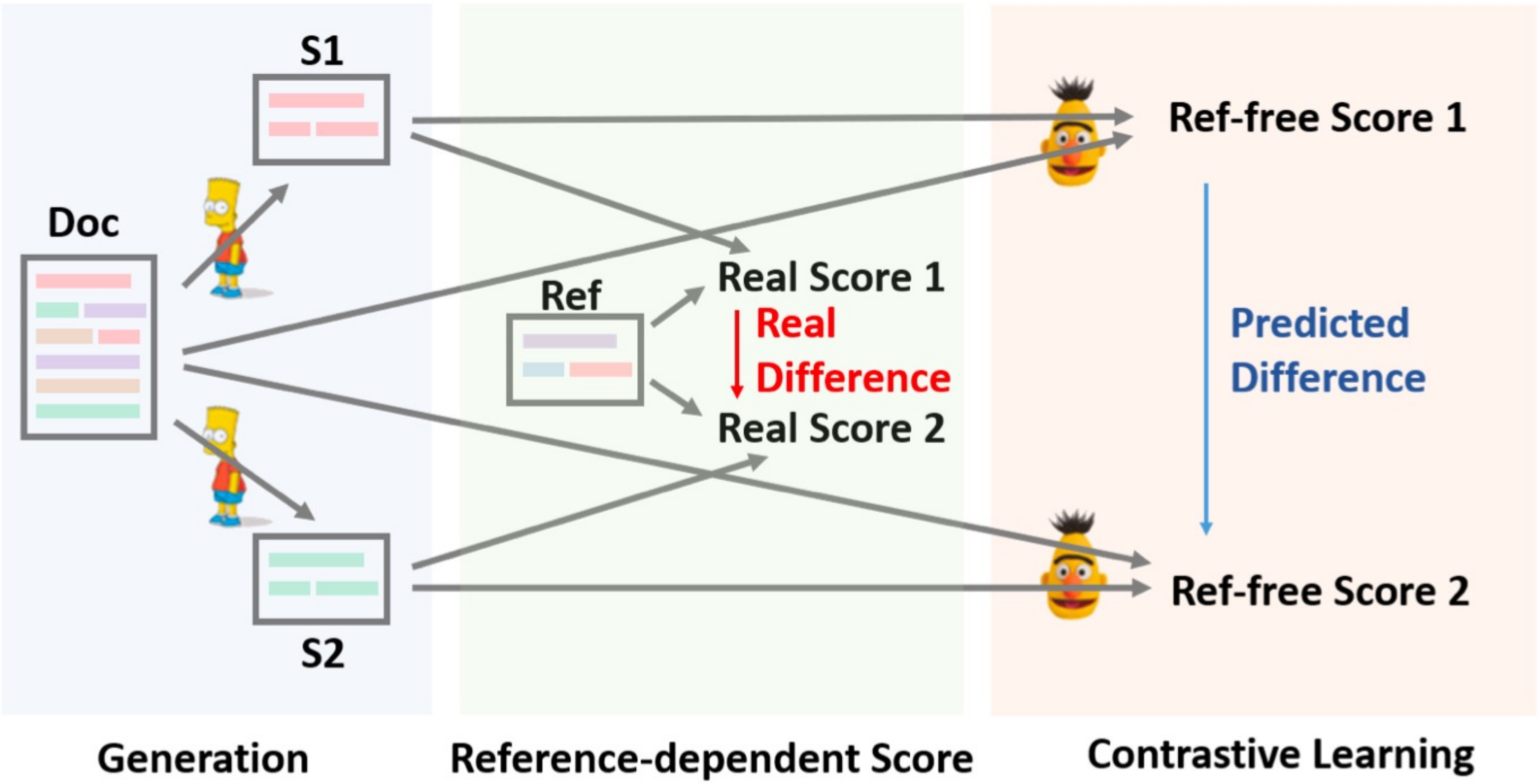
Contrastive Learning on Abstractive Summarization

- [2021ACL] SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization
- [2021]Alleviating Exposure Bias via Contrastive Learning for Abstractive Text Summarization
- [2021]Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization
- [2021]Sequence Level Contrastive Learning for Text Summarization
- [2021EMNLP]Topic-Aware Contrastive Learning for Abstractive Dialogue Summarization

SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization

- Train the summarization model as usual.
- **Generate multiple candidate summaries** during generation with diverse beam search. **(intuition: find the best one)**
- Train an **evaluation model** to rank the generated candidates with **contrastive learning**.
- The final output summary is the candidate with the highest score.

SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization



SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization

- Define pos&neg

Anchor Text	document
Positive Text	reference summary/generated summary (rank by ROUGE/Human)
Negative Text	generated summary

- Extract representations

  RoBERTa, cosine similarity

- Loss

SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization

- Loss

Generated summary

Reference summary

$$L = \sum_i \max(0, h(D, \boxed{\tilde{S}_i}) - h(D, \boxed{\hat{S}}))$$
$$+ \sum_i \sum_{j>i} \max(0, h(D, \tilde{S}_j) - h(D, \tilde{S}_i) + \lambda_{ij})$$

- $\tilde{S}_1, \cdots, \tilde{S}_n$ is descending sorted by evaluation metric (Rouge)
- $\lambda_{ij} = (j - i) * \lambda$ the corresponding margin

SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization

	CNNDM			XSum		
Model	R-1	R-2	R-L	R-1	R-2	R-L
BART	44.16	21.28	40.90	45.14	22.27	37.25
Pegasus	44.17	21.47	41.11	47.21	24.56	39.25
Prophet	44.20	21.17	41.30	-	-	-
GSum	45.94	22.32	42.48	45.40	21.89	36.67
Origin	44.39	21.21	41.28	47.10	24.53	39.23
Min	33.17	11.67	30.77	40.97	19.18	33.68
Max	54.36	28.73	50.77	52.45	28.28	43.36
Random	43.98	20.06	40.94	46.72	23.64	38.55
SimCLS	**46.67**	**22.15**	**43.54**	**47.61**	**24.57**	**39.44**

Alleviating Exposure Bias via Contrastive Learning for Abstractive Text Summarization

- **Intuition:** solving the discrepancy between training and inference

- **Silver summary**: the generated summary without beam search.
- Use contrastive learning as an auxiliary task during general training.

Alleviating Exposure Bias via Contrastive Learning for Abstractive Text Summarization

- Define pos&neg

Anchor Text	source document
Positive Text	gold summary
Negative Text	silver summary

- Extract representations

  PEGASUS

- Loss

Alleviating Exposure Bias via Contrastive Learning for Abstractive Text Summarization

- Loss

Pos score: $$S(Y|X) = \frac{1}{n^{\beta}} \sum_{i=1}^{n} \boxed{f(y_i|X, y_{<i})}$$

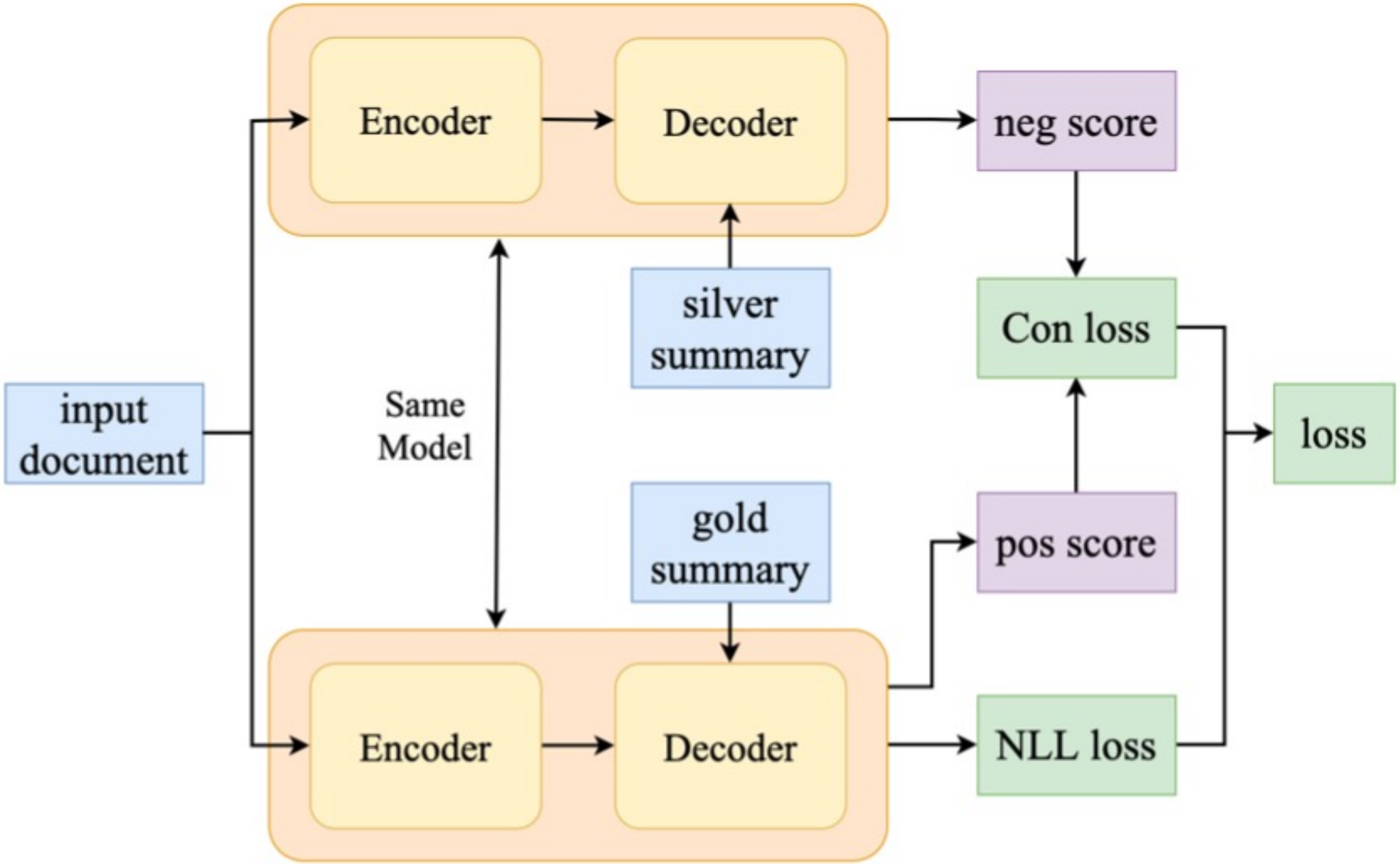
The predicted log-likelihood

Neg score: $$S(\hat{Y}|X) = \frac{1}{m^{\beta}} \sum_{i=1}^{m} f(\hat{y}_i|X, \widehat{y_{<i}})$$

$$L_{con} = \max(0, S(\hat{Y}|X) - S(Y|X) + \gamma)$$

$$L = L_{con} + L_{nll}$$

Alleviating Exposure Bias via Contrastive Learning for Abstractive Text Summarization



Alleviating Exposure Bias via Contrastive Learning for Abstractive Text Summarization

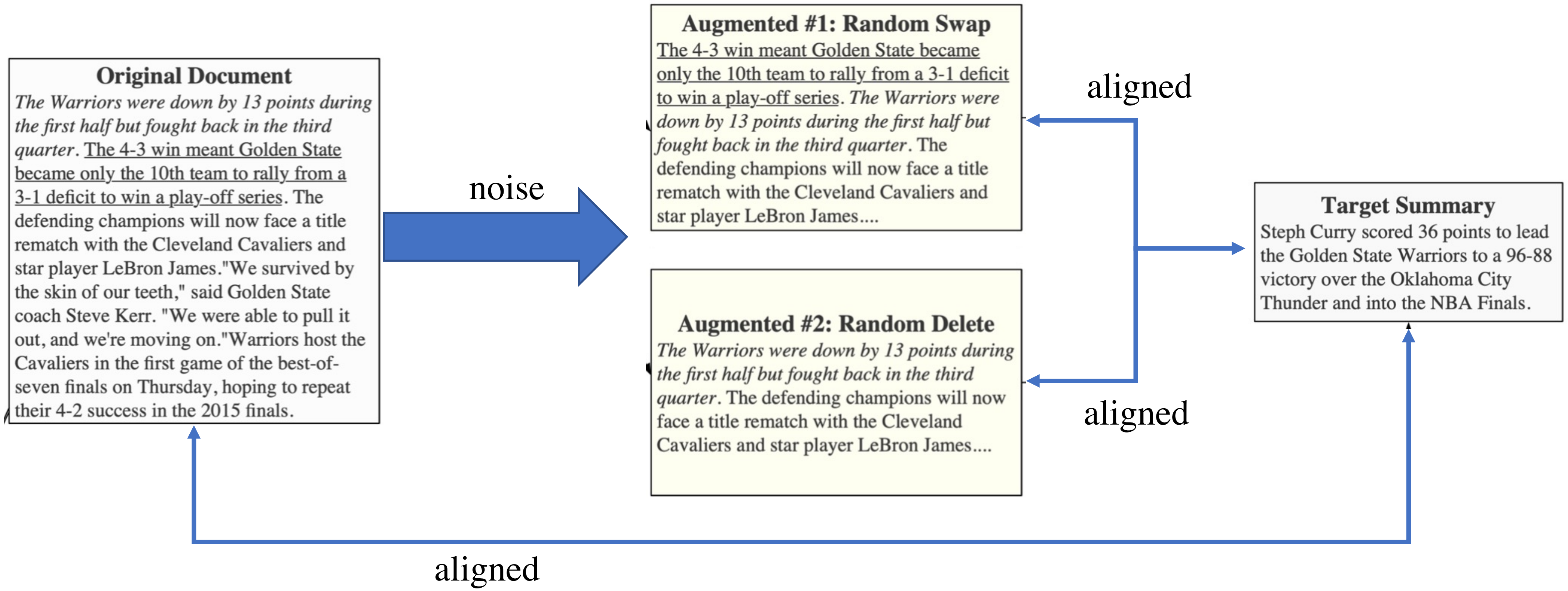
	CNNDM			**XSum**		
Model	R-1	R-2	R-L	R-1	R-2	R-L
BERTSUM	41.72	19.39	38.76	38.76	16.33	31.15
MASS	42.12	19.50	39.01	39.75	17.24	31.95
BART	44.16	21.28	40.90	45.14	22.27	37.25
ConSum	**44.53**	**21.54**	**41.57**	**47.34**	**24.67**	**39.40**

Contrastive Learning on Abstractive Summarization

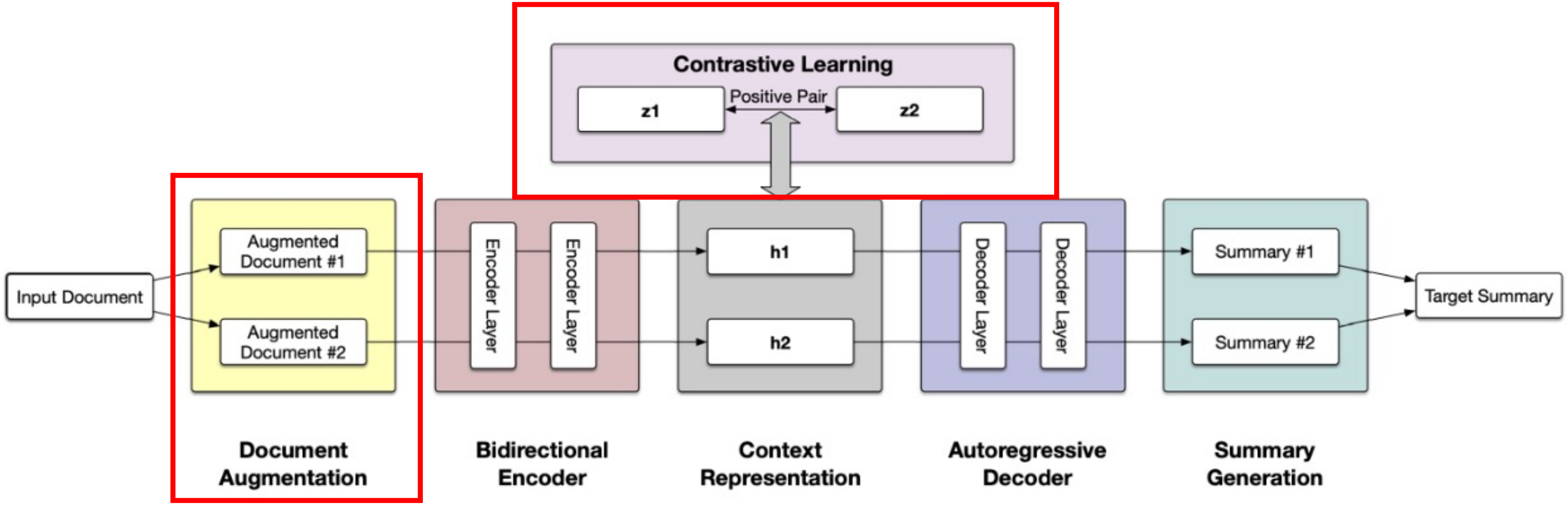
- [2021ACL] SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization
- [2021]Alleviating Exposure Bias via Contrastive Learning for Abstractive Text Summarization
- [2021]Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization
- [2021]Sequence Level Contrastive Learning for Text Summarization
- [2021EMNLP]Topic-Aware Contrastive Learning for Abstractive Dialogue Summarization

Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization

- Enhance the robustness of the model on noisy input documents



Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization



sentence-level document augmentation:

Random Insertion;

Random Swap;

Random Deletion;

~~Document Rotation~~

Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization

- Define pos/neg pairs

  Pos: if and only if two instances are from the same original input document.

  Neg: two different document
- Extract representations

  The final hidden vector of the first input token of BART's encoder.
- Loss

$$l(i,j) = -\log \frac{\exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2K} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization

	CNNDM			XSUM		
	R-1	R-2	R-L	R-1	R-2	R-L
Lead-3	40.07	17.68	36.33	16.30	1.60	11.95
BERTSUM	42.13	19.60	39.18	38.81	16.50	31.27
PGNet	36.44	15.66	33.42	29.70	9.21	23.24
BART	44.16	21.28	40.90	45.14	22.27	37.25
PEGASUS	44.17	21.47	41.11	47.21	24.56	39.25
Distil-BART	41.23	19.38	38.11	44.41	21.40	36.50
ESACL	44.24	21.06	41.20	44.64	21.62	36.73

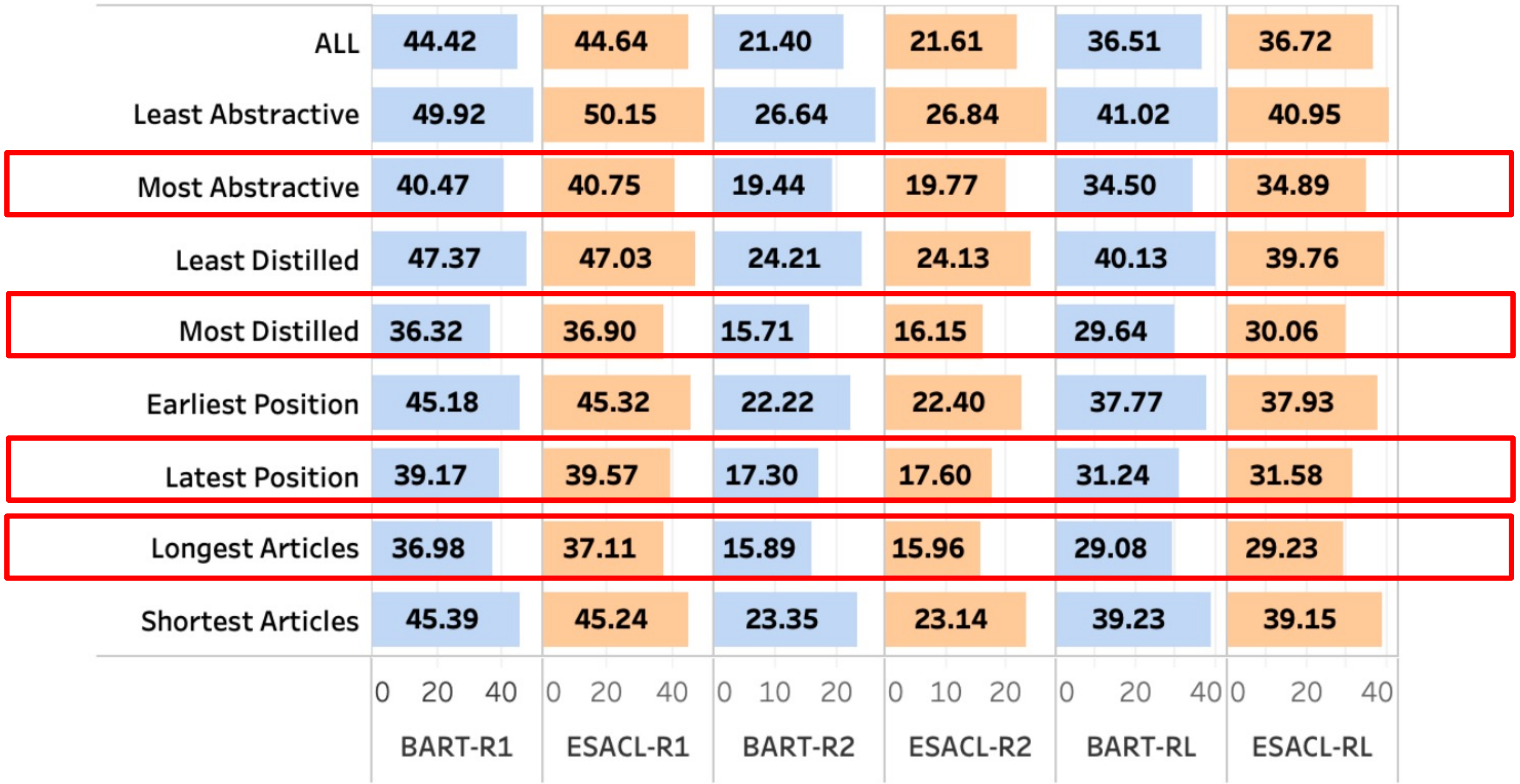
Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization

	CNNDM			XSUM		
	R-1	R-2	R-L	R-1	R-2	R-L
Lead-3	40.07	17.68	36.33	16.30	1.60	11.95
BERTSUM	42.13	19.60	39.18	38.81	16.50	31.27
PGNet	36.44	15.66	33.42	29.70	9.21	23.24
BART	44.16	21.28	40.90	45.14	22.27	37.25
PEGASUS	44.17	**21.47**	41.11	**47.21**	**24.56**	**39.25**
Distil-BART	41.23	19.38	38.11	44.41	21.40	36.50
ESACL	**44.24**	21.06	**41.20**	44.64	21.62	36.73

Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization

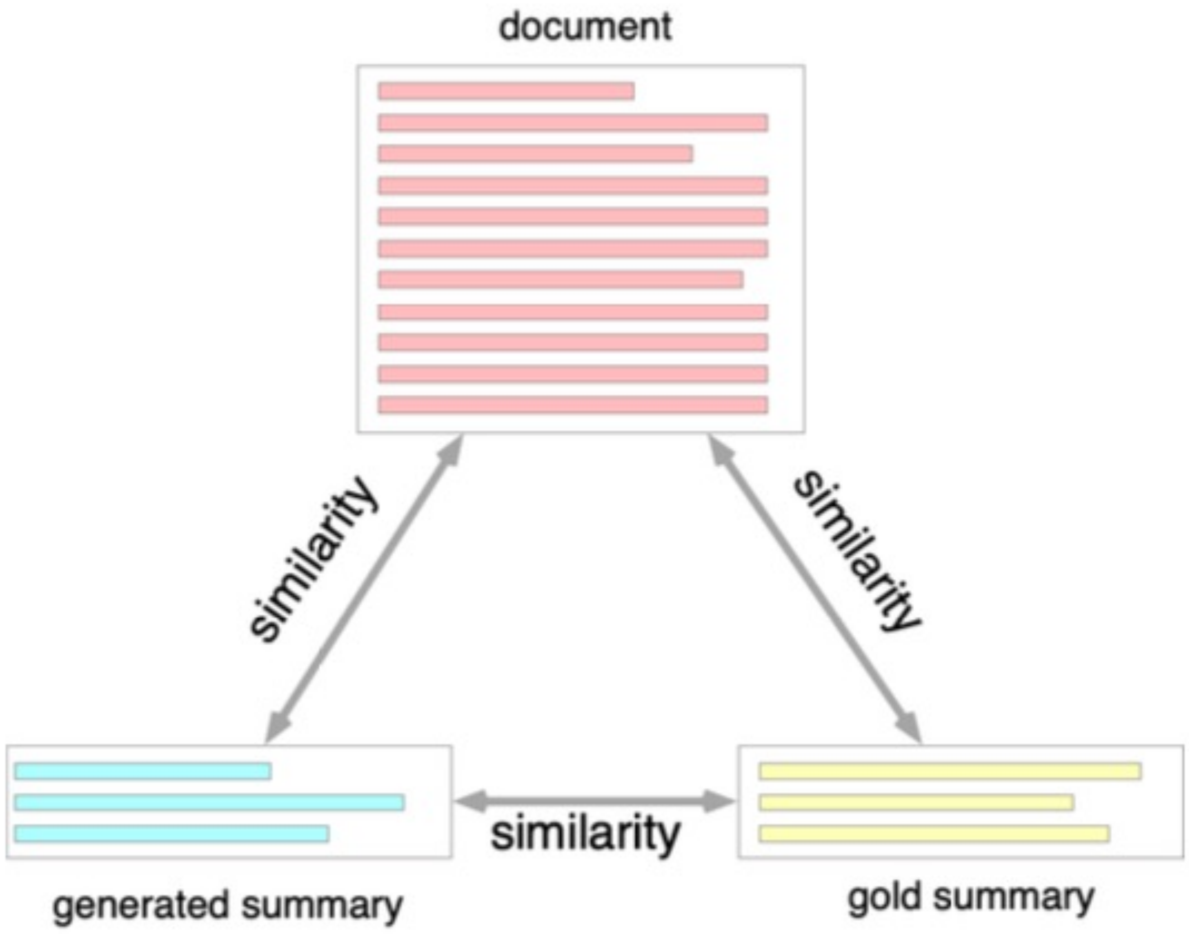
	CNNDM			XSUM		
	R-1	R-2	R-L	R-1	R-2	R-L
Lead-3	40.07	17.68	36.33	16.30	1.60	11.95
BERTSUM	42.13	19.60	39.18	38.81	16.50	31.27
PGNet	36.44	15.66	33.42	29.70	9.21	23.24
BART	44.16	21.28	40.90	45.14	22.27	37.25
PEGASUS	44.17	21.47	41.11	47.21	24.56	39.25
Distil-BART	41.23	19.38	38.11	44.41	21.40	36.50
ESACL	44.24	21.06	41.20	44.64	21.62	36.73

Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization



- Abstractiveness
- Distillation
- Position
- Length

Sequence Level Contrastive Learning for Text Summarization



Sequence Level Contrastive Learning for Text Summarization

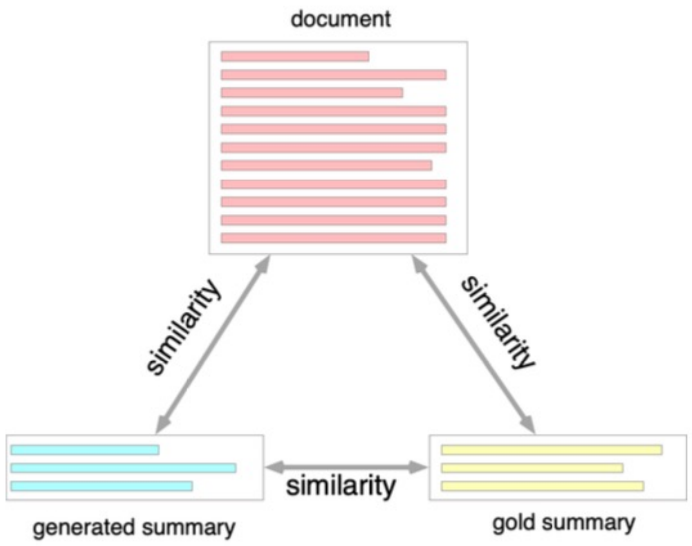
- Define pos&neg

  POS: a document, its gold summary and its generated summary

- Extract representations

  Encoder & Decoder

- Loss



$$\mathcal{L} = \mathcal{L}^{\text{NLL}} + \lambda_{x-y}\mathcal{L}^{\text{E}}_{\text{sim}}(X, Y) + \lambda_{x-\hat{y}}\mathcal{L}^{\text{E}}_{\text{sim}}(X, \hat{Y}) + \lambda_{y-\hat{y}}\mathcal{L}^{\text{E}}_{\text{sim}}(Y, \hat{Y}) + \lambda^{\text{D}}_{y-\hat{y}}\mathcal{L}^{\text{D}}_{\text{sim}}(Y, \hat{Y})$$

Sequence Level Contrastive Learning for Text Summarization

	CNNDM			XSUM		
	R-1	R-2	R-L	R-1	R-2	R-L
Lead-3	40.07	17.68	36.33	16.30	1.60	11.95
BERTSUM	42.13	19.60	39.18	38.81	16.50	31.27
PGNet	36.44	15.66	33.42	29.70	9.21	23.24
BART	44.16	21.28	40.90	45.14	22.27	37.25
PEGASUS	44.17	21.47	41.11	47.21	24.56	39.25
SeqCo(x-y)	44.66	21.57	41.38	**45.65**	**22.41**	**37.04**
SeqCo(x-y’)	44.94	**21.82**	41.68	45.60	22.36	36.94
SeqCo(y-y’)	**45.02**	21.80	**41.75**	45.52	22.24	36.90

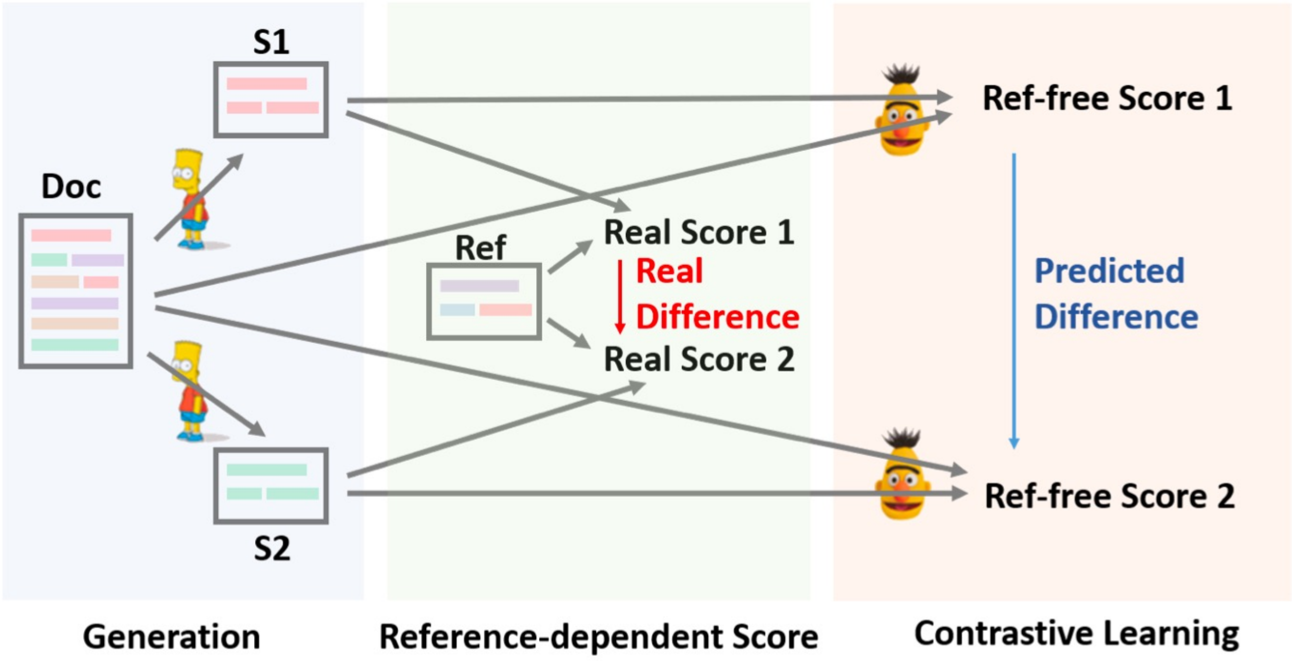
Comparisons between above methods

SimCLS

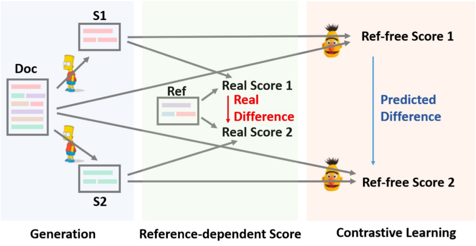
ConSum

ESACL

SeqCo



Comparisons between above methods

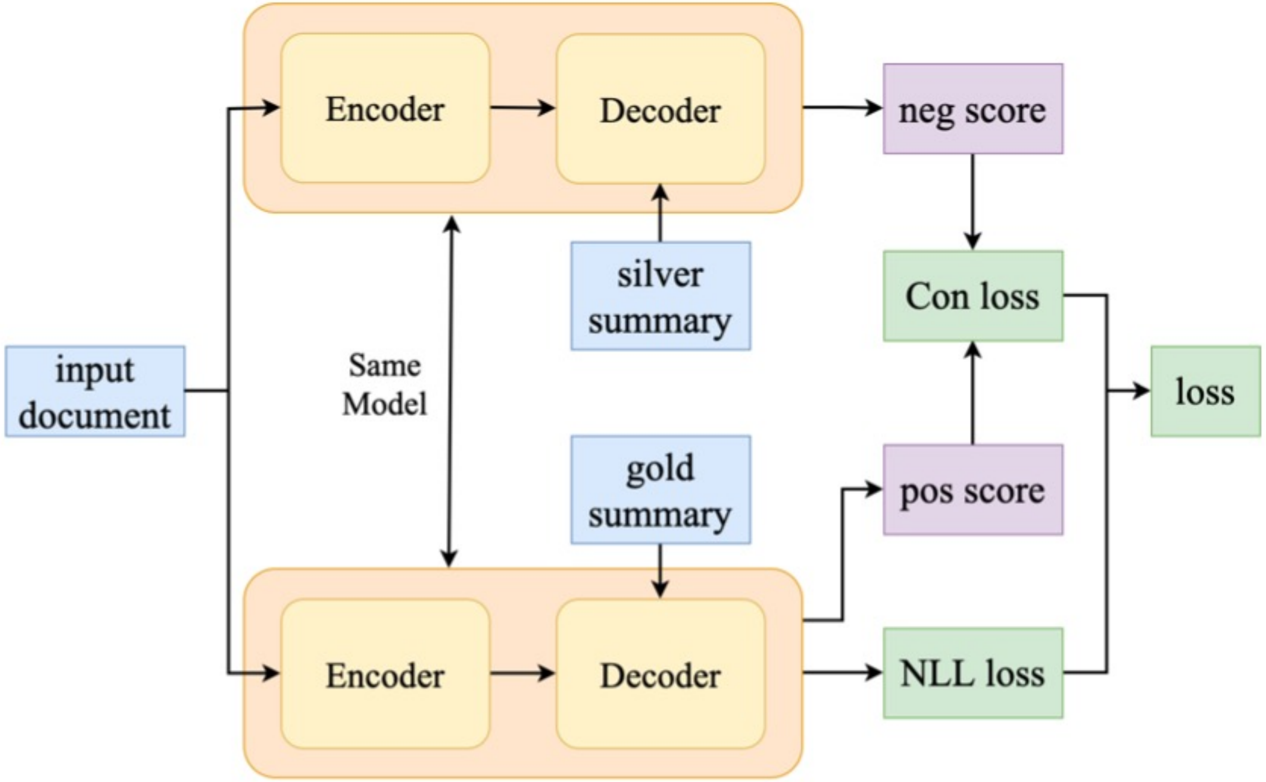


SimCLS

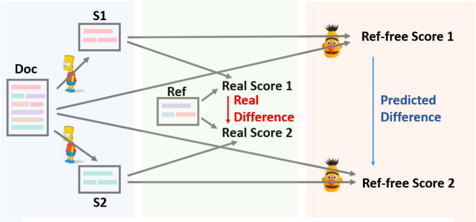
ConSum

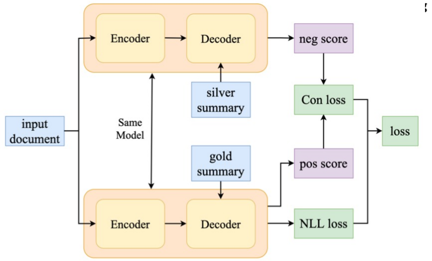
ESACL

SeqCo



Comparisons between above methods



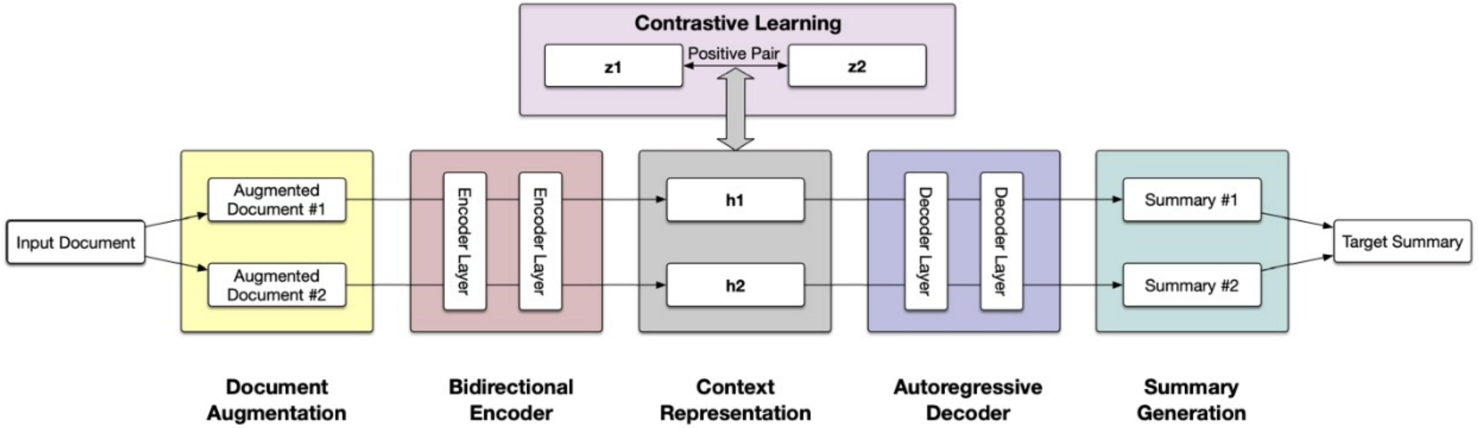


SimCLS

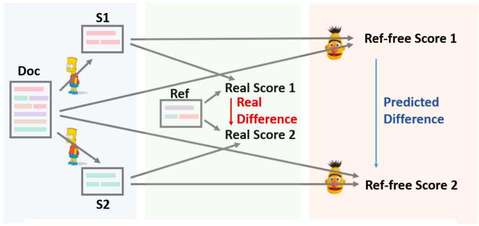
ConSum

ESACL

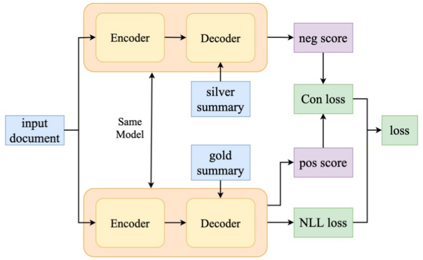
SeqCo



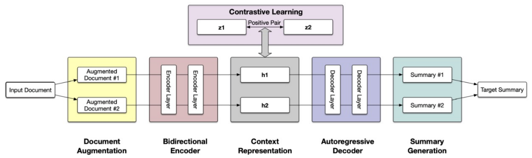
Comparisons between above methods



SimCLS

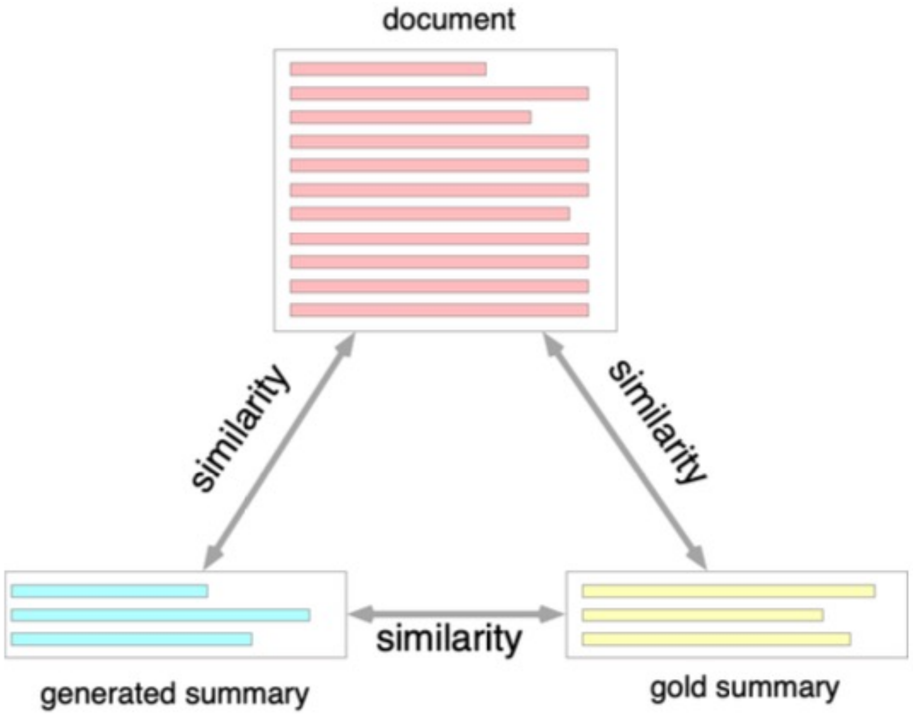


ConSum

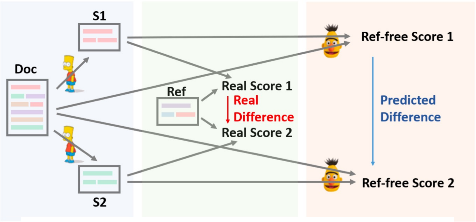


ESACL

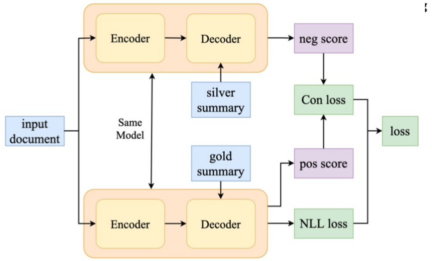
SeqCo



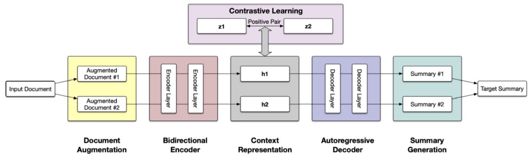
Comparisons between above methods



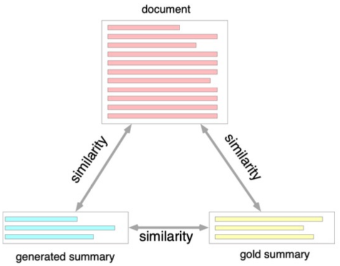
SimCLS



ConSum



ESACL



SeqCo

Q
U
I
Z

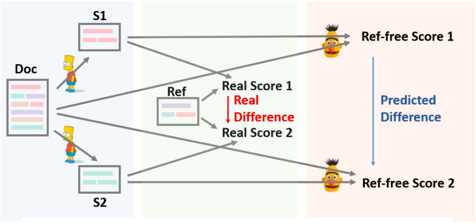
A lack of representing similarities with NLL

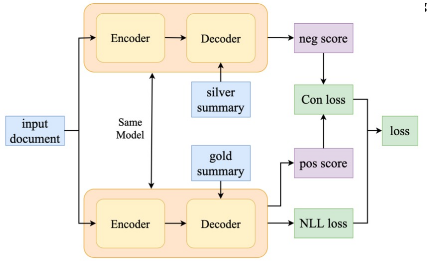
Deal with noisy document input

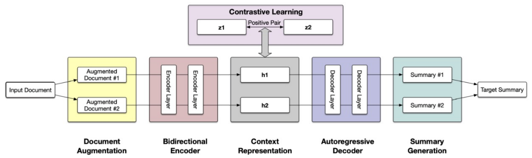
Narrow down the gap between NLL and Rouge

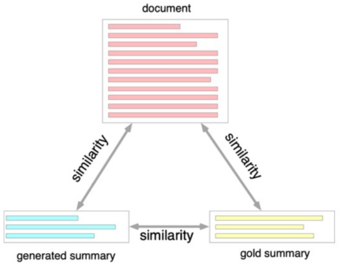
Narrow down the gap between traning and testing

Comparisons between above methods









SimCLS

ConSum

ESACL

SeqCo

A lack of representing similarities with NLL

Deal with noisy document input

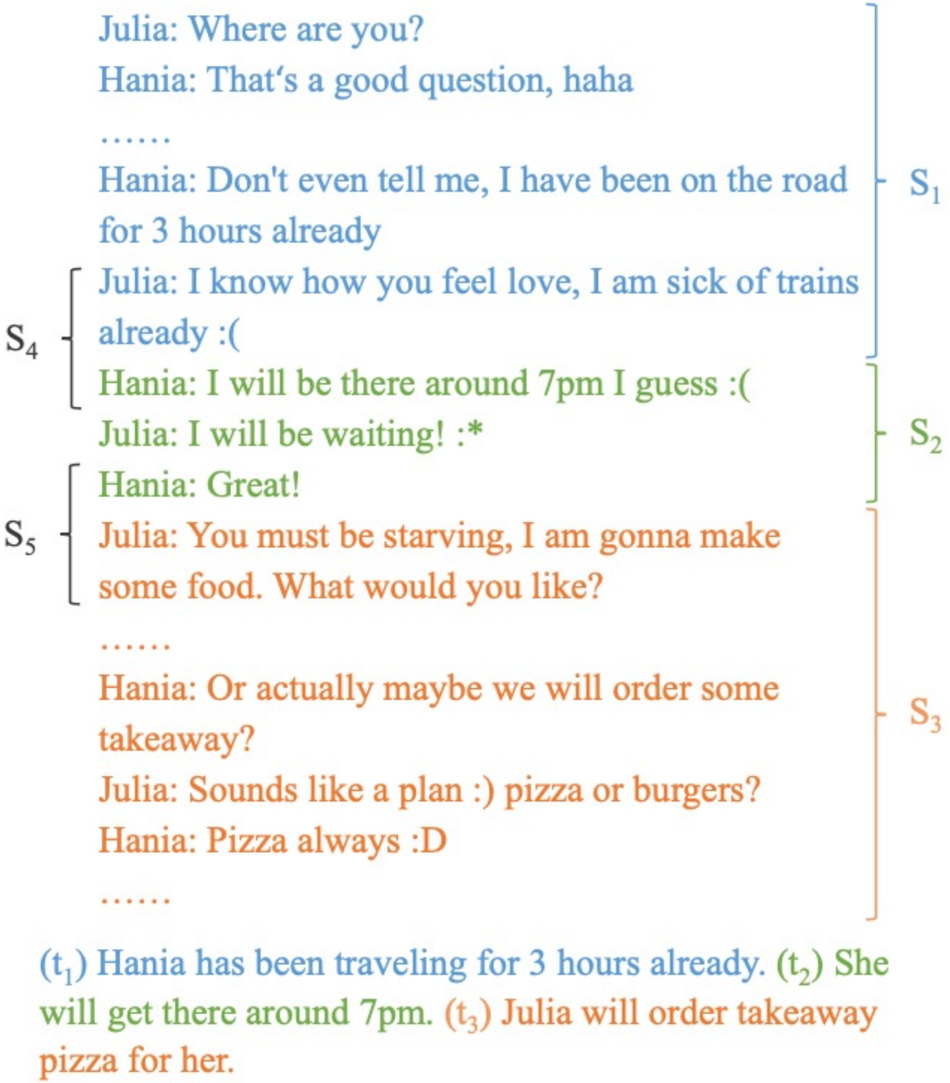
Narrow down the gap between NLL and Rouge

Narrow down the gap between traning and testing

Comparisons between above methods

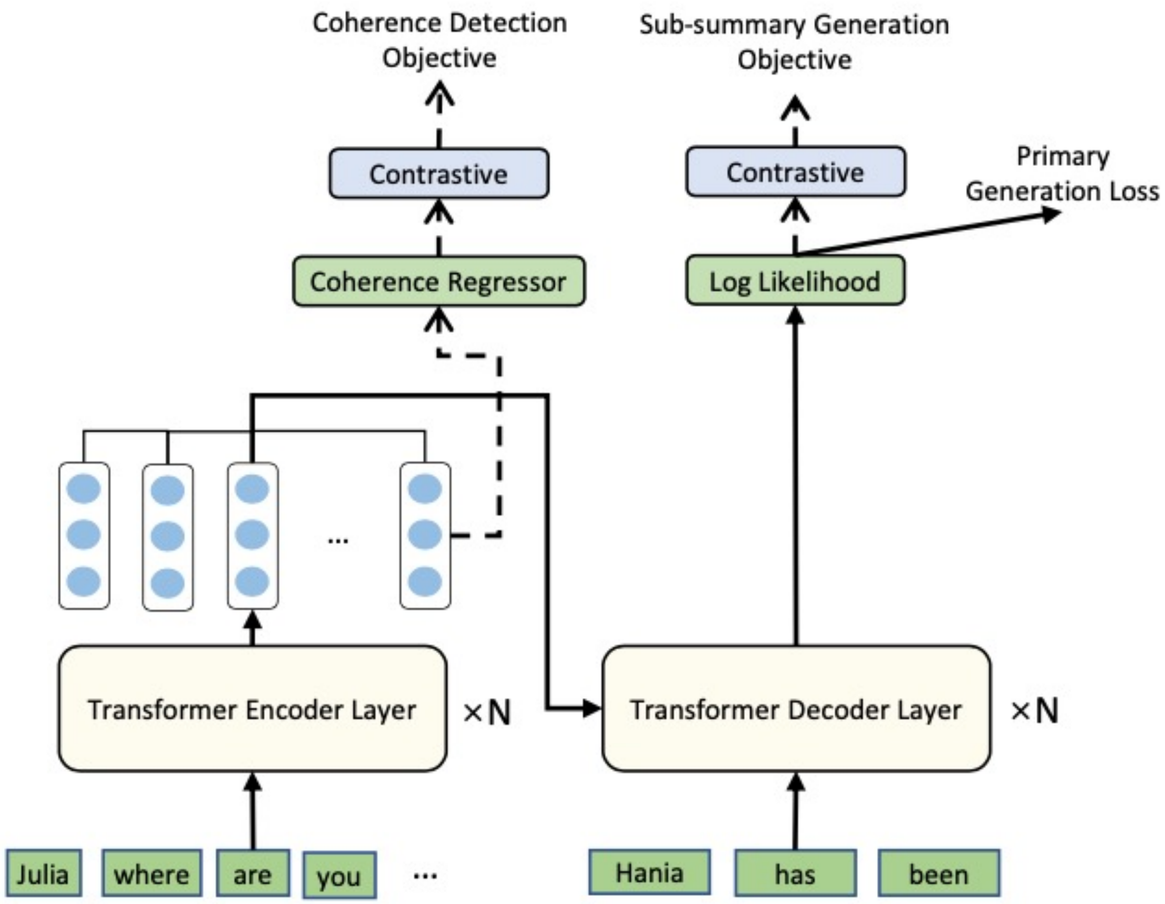
	CNNDM			XSUM		
	R-1	R-2	R-L	R-1	R-2	R-L
Lead-3	40.07	17.68	36.33	16.30	1.60	11.95
BERTSUM	42.13	19.60	39.18	38.81	16.50	31.27
PGNet	36.44	15.66	33.42	29.70	9.21	23.24
Distil-BART	41.23	19.38	38.11	44.41	21.40	36.50
BART	44.16	21.28	40.90	45.14	22.27	37.25
PEGASUS	44.17	21.47	41.11	47.21	24.56	39.25
SimCLS	**46.67**	**22.15**	**43.54**	**47.61**	24.57	**39.44**
ConSum	44.53	21.54	41.57	47.34	**24.67**	39.40
ESACL	44.24	21.06	41.20	44.64	21.62	36.73
SeqCo	45.02	21.80	41.75	45.65	22.41	37.04

Topic-Aware Contrastive Learning for Abstractive Dialogue Summarization



- coherence detection
- sub-summary generation objectives

Topic-Aware Contrastive Learning for Abstractive Dialogue Summarization



- coherence detection
- sub-summary generation objectives

Topic-Aware Contrastive Learning for Abstractive Dialogue Summarization

Coherence Detection Objective

- Define pos&neg

    pos: a snippet

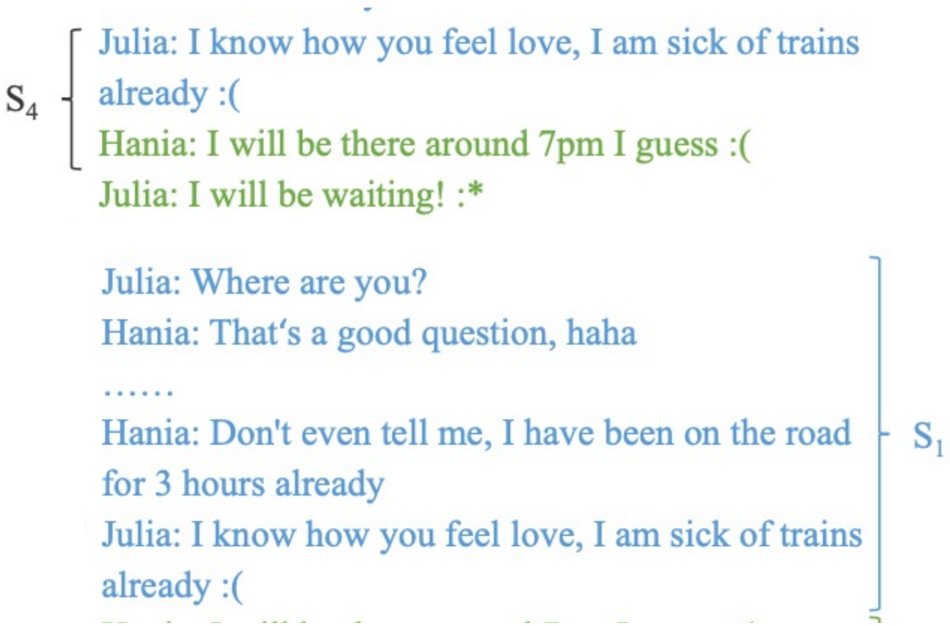
    neg: a shuffled snippet

- Extract representations

    last hidden state of Encoder

    + linear transformation

- Loss



$$[co(\mathcal{S}_k^{\mathcal{D}}), co(\widetilde{\mathcal{S}_k^{\mathcal{D}}})] = softmax([y_{\mathcal{S}_k^{\mathcal{D}}}, y_{\widetilde{\mathcal{S}_k^{\mathcal{D}}}}])$$

$$\mathcal{L}_{co}^{\mathcal{D}} = \frac{1}{N_{co}} \sum_{n=1}^{N_{co}} \max(0, \delta_{co} - (co(S_{k,n}^{\mathcal{D}}) - co(\widetilde{S_{k,n}^{\mathcal{D}}})))$$

Topic-Aware Contrastive Learning for Abstractive Dialogue Summarization

Sub-summary Generation Objective

- Define pos&neg

  pos: the most related snippet (compared with sub-summary)

  neg: a random snippet from the rest

- Extract representations

$$\mathcal{L}_{pos}^{t_i} = -\log(\prod_{j=1}^{|t_i|} p(t_j^i | t_{1:j-1}^i, \mathcal{S}_{\text{pos}}^i; \theta))$$

$$\mathcal{L}_{neg}^{t_i} = -\log(\prod_{j=1}^{|t_i|} p(t_j^i | t_{1:j-1}^i, \mathcal{S}_{\text{neg}}^i; \theta))$$

- Loss

$$[su(S_{\text{pos}}^i), su(S_{\text{neg}}^i)] = softmax([\mathcal{L}_{pos}^{t_i}, \mathcal{L}_{neg}^{t_i}])$$

$$\mathcal{L}_{su}^{\mathcal{D},T_{\mathcal{D}}} = \frac{1}{N_{su}} \sum_{n=1}^{N_{su}} \max(0, \delta_{su} - (su(S_{\text{neg}}^n) - su(S_{\text{pos}}^n)))$$

Topic-Aware Contrastive Learning for Abstractive Dialogue Summarization

Model	R-1	R-2	R-L	BERTS
∗Lead3	31.4	8.7	29.4	-
∗PTGen	40.1	15.3	36.6	-
∗DynamicConv + GPT-2	41.8	16.4	37.6	-
∗FastAbs-RL	42.0	18.1	39.2	-
∗DynamicConv + News	45.4	20.7	41.5	-
Multiview BART	53.9	28.4	44.4	53.6
∗$\text{BART}_{\text{BASE}}$	46.1	22.3	36.4	44.8
∗BART	52.6	27.0	42.1	52.1
∗BART_{ORI}	52.6	27.2	42.7	52.3
$\text{CONDIGSUM}_{\text{BASE}}$	48.1	24.0	39.2	48.0
CONDIGSUM	**54.3**	**29.3**	**45.2**	**54.0**
w/o Sub-summary	53.8	28.3	44.1	53.5
w/o Coherence	53.9	28.6	44.2	53.5

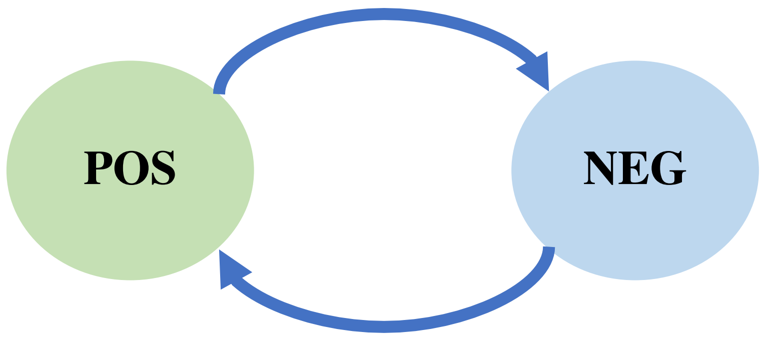
Comparison & Conclusion

- Motivation
  - ✓ Summary quality on a specific aspect
  - ✓ Model robustness
  - ✓ Exposure bias

Together?

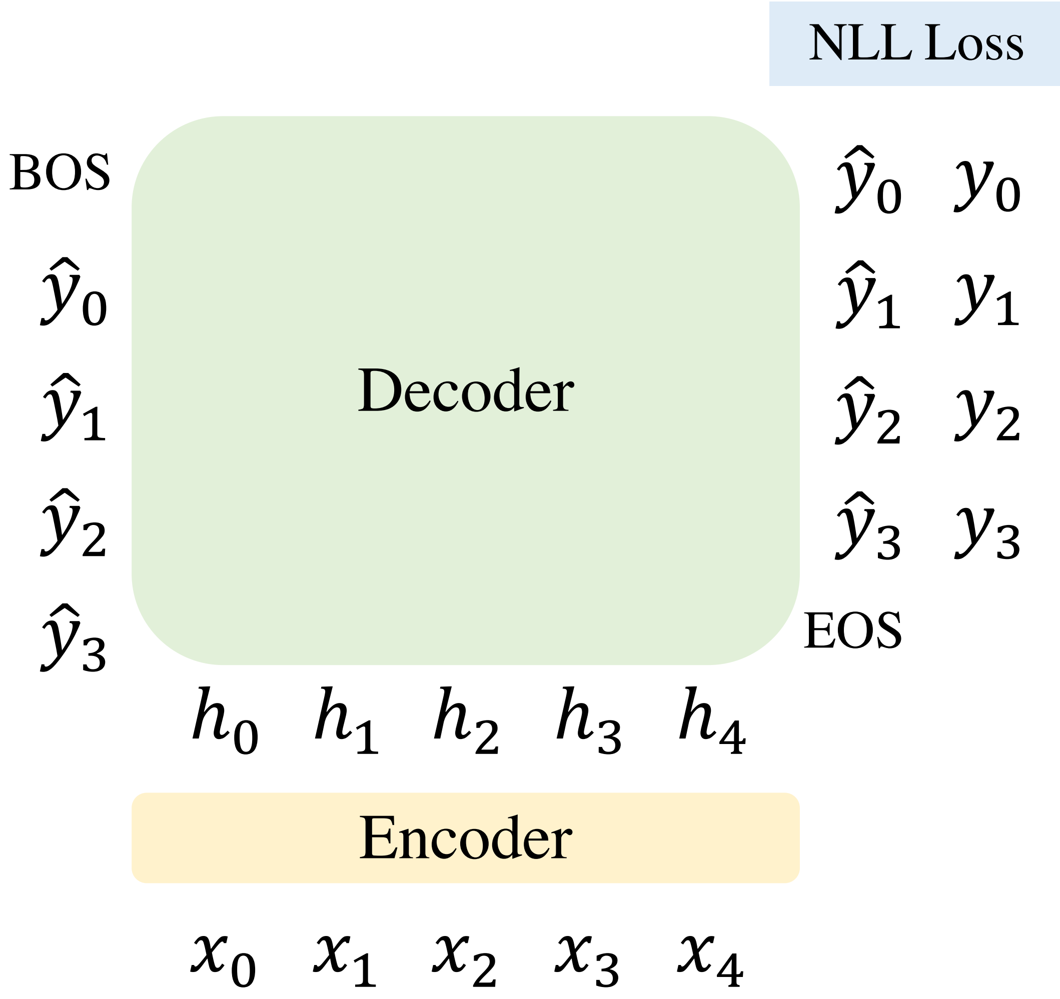
Comparison & Conclusion

- Construct Positive Pairs and Negative Pairs (Within a batch)
  - ✓ Word level, Sentence level, Discourse level
  - ✓ Document-only, Summary-only, Document-Summary
  - ✓ Insertion, Replacement, Delete, Rotate



Comparison & Conclusion

- Extract Representations
  - ✓ Encoder hidden states: the first, the last, all
  - ✓ Decoder hidden states
  - ✓ + linear/non-linear/multi-head attention
  - ✓ Decoder loss



Comparison & Conclusion

- ➢ Similarity:
  - ✓ Cosine similarity
  - ✓ Softmax between (pos, neg)
- ➢ Loss
  - ✓ Weighted-sum with NLL
  - ✓ Alternating Update Strategy

Algorithm 2 Alternating Updating Strategy

Input: A batch of dialogue-summary instances $\mathcal{B}$

Coherence Task

1: $\mathcal{L}_{co}^{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{\langle \mathcal{D}, T_{\mathcal{D}} \rangle \in \mathcal{B}} \mathcal{L}_{co}^{\mathcal{D}}$

2: $\theta \leftarrow \theta - \alpha w_{co} \frac{\partial L_{co}^{\mathcal{B}}}{\partial \theta}$

Sub-summary Task

3: $\mathcal{L}_{su}^{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{\langle \mathcal{D}, T_{\mathcal{D}} \rangle \in \mathcal{B}} \mathcal{L}_{su}^{\mathcal{D}, T_{\mathcal{D}}}$

4: $\theta \leftarrow \theta - \alpha w_{su} \frac{\partial \mathcal{L}_{su}^{\mathcal{B}}}{\partial \theta}$

Main Task

5: $\mathcal{L}_{main}^{\mathcal{B}} = -\frac{1}{|\mathcal{B}|} \sum_{\langle \mathcal{D}, T_{\mathcal{D}} \rangle \in \mathcal{B}} \mathcal{L}^{\mathcal{D}, T_{\mathcal{D}}}$

6: $\theta \leftarrow \theta - \alpha w_{main} \frac{\partial \mathcal{L}_{main}^{\mathcal{B}}}{\partial \theta}$

$$\mathcal{L} = \alpha \mathcal{L}_{\text{cl}} + (1 - \alpha) \mathcal{L}_{\text{generate}}$$

Thank you!