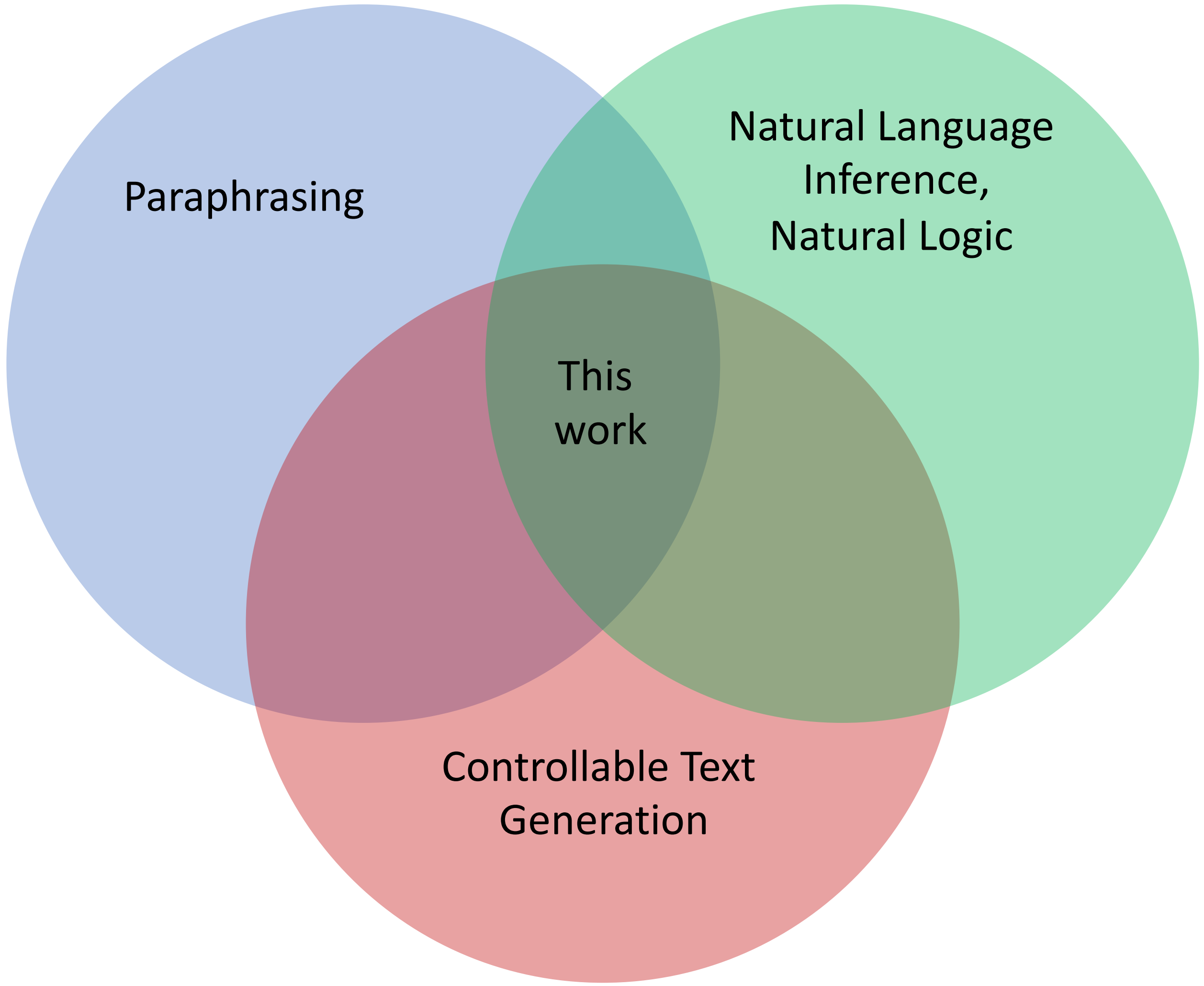
1

2

Entailment Relation Aware Paraphrase Generation

Abhilasha Sancheti[1,2], Balaji Vasan Srinivasan[2], Rachel Rudinger[1]


Paraphrasing
Natural Language
Inference,
Natural Logic
This
work
Controllable Text
Generation

Paraphrase Generation (Entailment-Unaware)

Input: A young girl is looking at a women in a costume.

Paraphrase Generation (Entailment-Unaware)

Input: A young girl is looking at a women in a costume.

Output 1: A girl is looking at a person in a costume.

Paraphrase Generation (Entailment-Unaware)

Input: A young girl is looking at a women in a costume.

Output 1: A girl is looking at a person in a costume.

Output 2: A young girl looks at a women dressed up as a witch.

Paraphrase Generation (Entailment-Unaware)

Input: A young girl is looking at a women in a costume.

Output 1: A girl is looking at a person in a costume.

Output 2: A young girl looks at a women dressed up as a witch.

Output 3: A girl, who is little, looks at a women in a costume.

Paraphrase Generation (Entailment-Unaware)

Input: A young girl is looking at a women in a costume.

Output 1: A girl is looking at a person in a costume.	Forward Entailment (⊏)

Output 2: A young girl looks at a women dressed up as a witch.

Output 3: A girl, who is little, looks at a women in a costume.

Paraphrase Generation (Entailment-Unaware)

Input: A young girl is looking at a women in a costume.

Output 1: A girl is looking at a person in a costume.	Forward Entailment (⊏)
Output 2: A young girl looks at a women dressed up as a witch.	Reverse Entailment (⊐)

Output 3: A girl, who is little, looks at a women in a costume.

Paraphrase Generation (Entailment-Unaware)

Input: A young girl is looking at a women in a costume.

Output 1: A girl is looking at a person in a costume.	Forward Entailment ($\sqsubset$)
Output 2: A young girl looks at a women dressed up as a witch.	Reverse Entailment ($\sqsupset$)
Output 3: A girl, who is little, looks at a women in a costume.	Equivalence ($\equiv$)

Entailment-Aware Paraphrase Generation

Desired entailment relation

Input: (A young girl is looking at a women in a costume., $\sqsubset$)

Entailment-Aware Paraphrase Generation

Desired entailment relation

Input: (A young girl is looking at a women in a costume., ⊏)

Output 1: A girl is looking at a person in a costume.	Forward Entailment (⊏)

Entailment-Aware Paraphrase Generation

Desired entailment relation

Input: (A young girl is looking at a women in a costume.,⊐)

Output 2: A young girl looks at a women dressed up as a witch.	Reverse Entailment (⊐)

Entailment-Aware Paraphrase Generation

Desired entailment relation

Input: (A young girl is looking at a women in a costume.,≡)

Output 3: A girl, who is little, looks at a women in a costume.	Equivalence (≡)

Entailment Relations (MacCartney, 2009)

Entailment Relations (MacCartney, 2009)

Forward Entailment	$X \sqsubset Y$:= If X is true then Y is true.

Entailment Relations (MacCartney, 2009)

Forward Entailment	X $\sqsubset$ Y := If X is true then Y is true.
Reverse Entailment	X $\sqsupset$ Y := If Y is true then X is true.

Entailment Relations (MacCartney, 2009)

Forward Entailment	$X \sqsubset Y$:= If X is true then Y is true.
Reverse Entailment	$X \sqsupset Y$:= If Y is true then X is true.
Equivalence	$X \equiv Y$:= X is true if and only if Y is true ($X \sqsubset Y$ and $X \sqsupset Y$).

Entailment Relations (MacCartney, 2009)

Forward Entailment	$X \sqsubset Y$:= If X is true then Y is true.
Reverse Entailment	$X \sqsupset Y$:= If Y is true then X is true.
Equivalence	$X \equiv Y$:= X is true if and only if Y is true ($X \sqsubset Y$ and $X \sqsupset Y$).
Neutral	If X is true then Y may be true or false (cannot determine).

Entailment Relations (MacCartney, 2009)

Forward Entailment	$X \sqsubset Y$:= If X is true then Y is true.
Reverse Entailment	$X \sqsupset Y$:= If Y is true then X is true.
Equivalence	$X \equiv Y$:= X is true if and only if Y is true ($X \sqsubset Y$ and $X \sqsupset Y$).
Neutral	If X is true then Y may be true or false (cannot determine).
Contradiction	If X is true then Y is false and if Y is true then X is false.

Entailment Relations (MacCartney, 2009)

Forward Entailment	$X \sqsubset Y$:= If X is true then Y is true.
Reverse Entailment	$X \sqsupset Y$:= If Y is true then X is true.
Equivalence	$X \equiv Y$:= X is true if and only if Y is true ($X \sqsubset Y$ and $X \sqsupset Y$).
Neutral	If X is true then Y may be true or false (cannot determine).
Contradiction	If X is true then Y is false and if Y is true then X is false.

Kind of Relations in Paraphrase Datasets

Relation	ParaNMT (Wieting et al., 2018)	ParaBank (Hu et al., 2019)
Equivalence (≡)	55%	73.3%
Forward Entailment (⊏)	20%	8.1%
Reverse Entailment (⊐)	8%	7.8%
Neutral	8.5%	4.6%
Contradiction	3%	1.9%
Other/Invalid	5.5%	4.3%

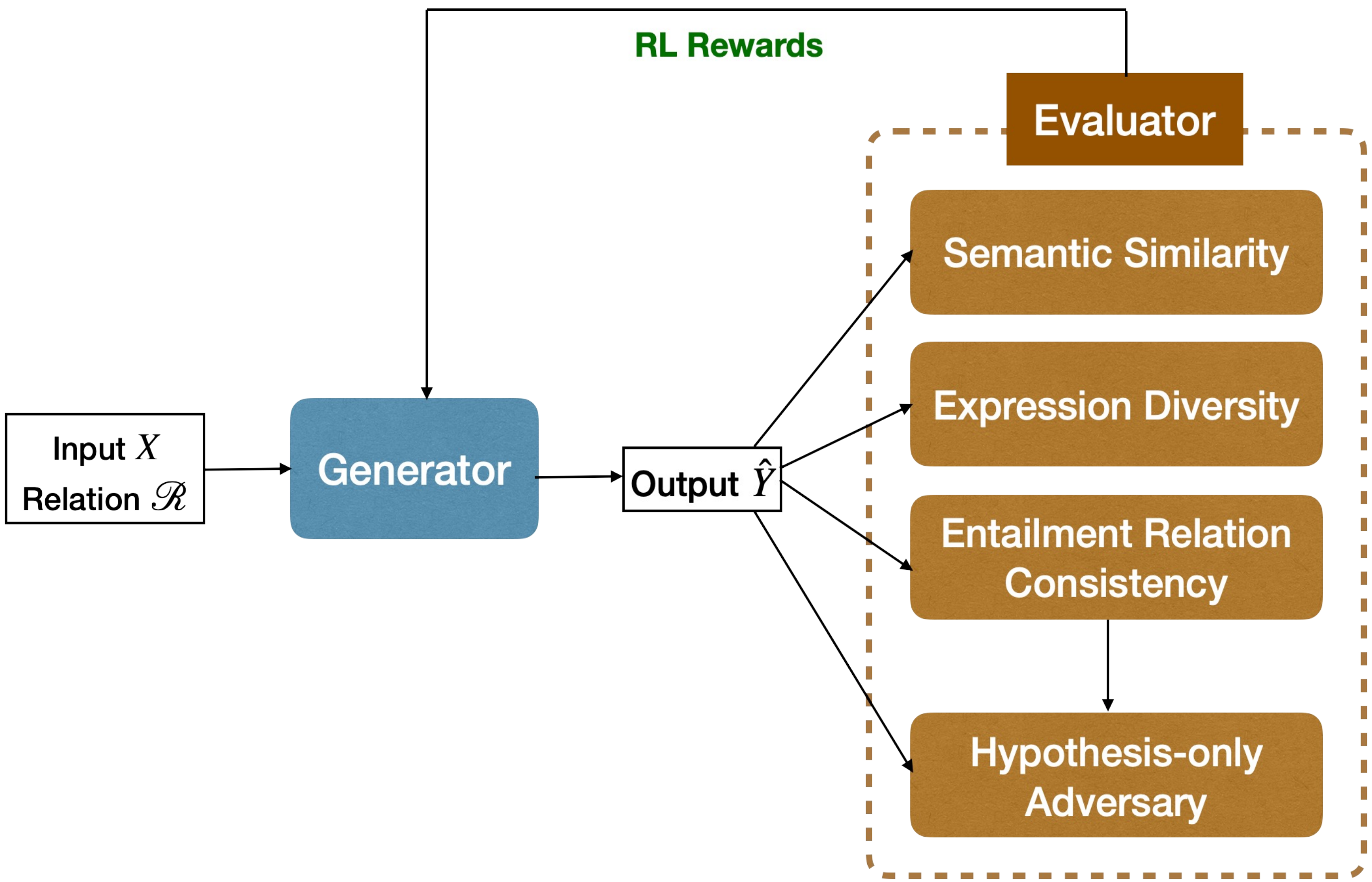
Why Entailment-Aware Paraphrasing?

- Equivalent (≡) paraphrases for rewriting in highly conservative, precise contexts.
    - Legal NLP, Medical NLP, paraphrastic data augmentation
- Forward Entailment (⊏) when details may be dropped
    - summarization, simplification, information retrieval
- Reverse Entailment (⊐) in creative contexts where new details may be introduced
    - storytelling, conversational AI, artistic expression or entertainment, information retrieval
- Paraphrastic dataset augmentation may use all three relations
    - E.g. Natural Language Inference

How to get Entailment Relation Labels?

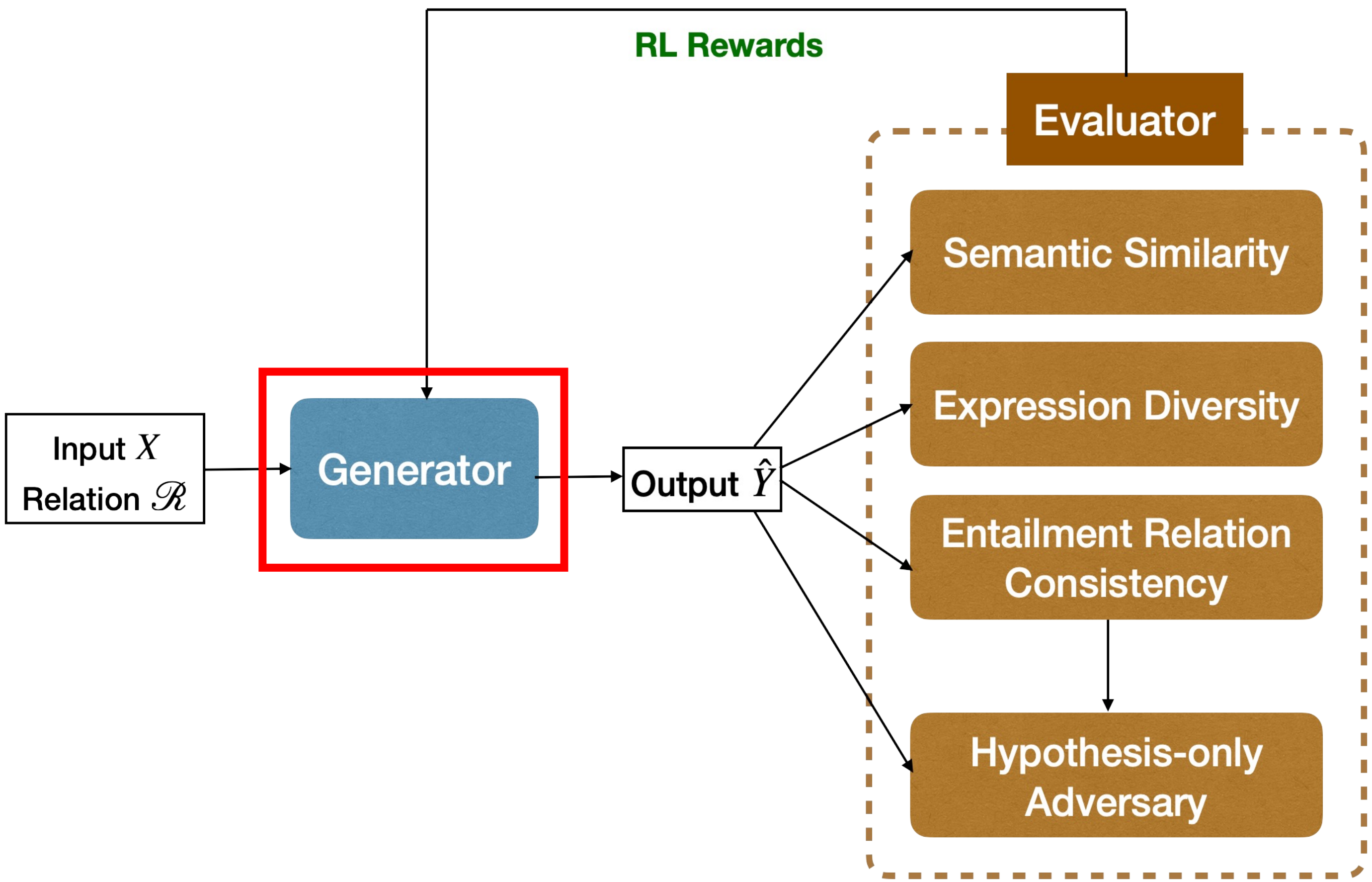
- Training a supervised entailment-aware paraphrasing system needs paraphrases labeled with entailment relations
  - Manual annotation is expensive
- Addressing data challenge in 3 ways
  - Recasting SICK (Marelli et al. 2014) dataset; using meaning preserving sentences and uni-directional entailment relations
  - Developing NLI-trained Entailment Relation Oracle to obtain weak-supervision for entailment labels
  - Developing **E**ntailment **R**elation **A**ware **P**araphraser that can trained using existing Paraphrase and NLI datasets

ERAP: Entailment Relation Aware Paraphraser



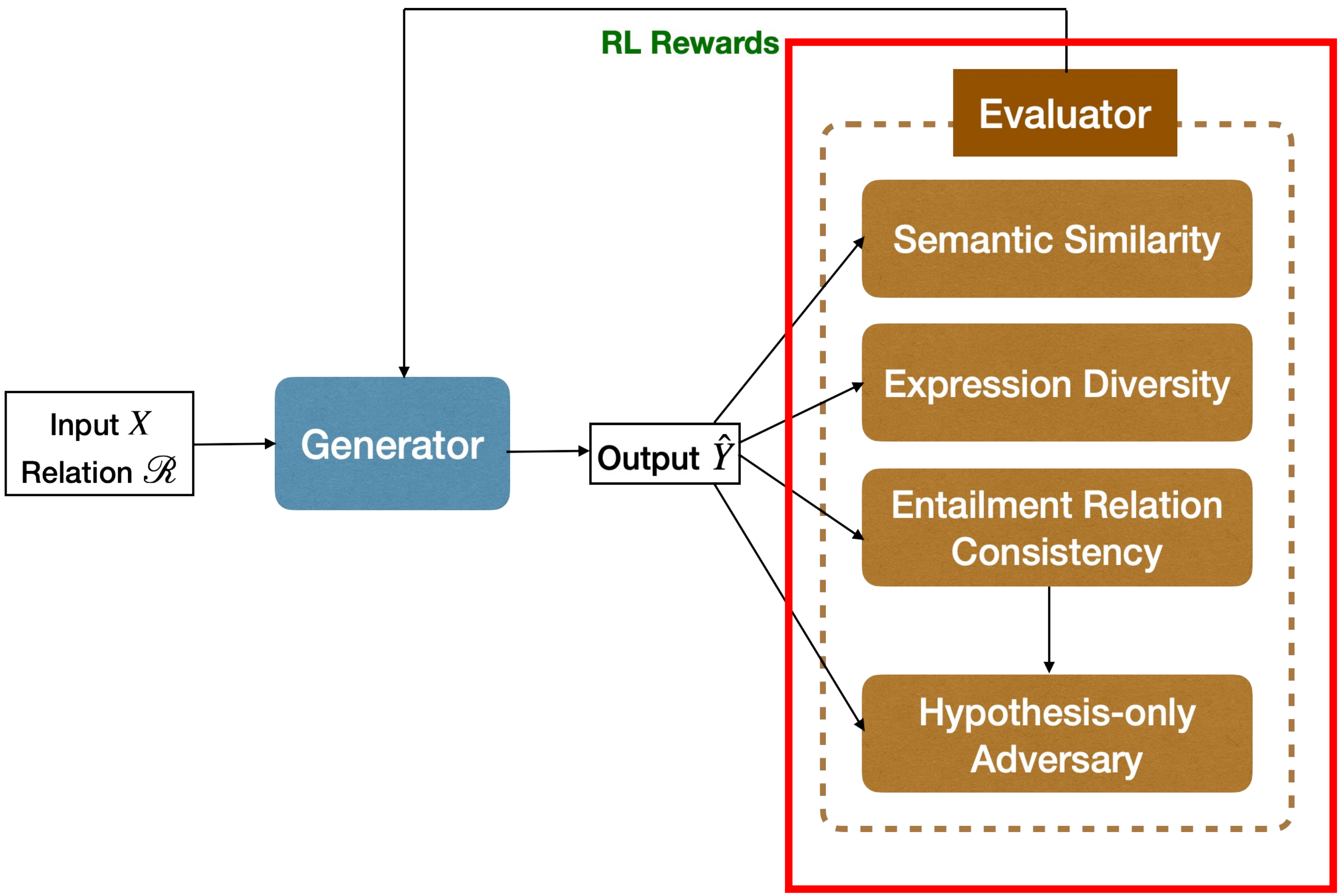
ERAP: Entailment Relation Aware Paraphraser

Generator:
Takes in a sentence and an entailment relation for conditioning the generation.



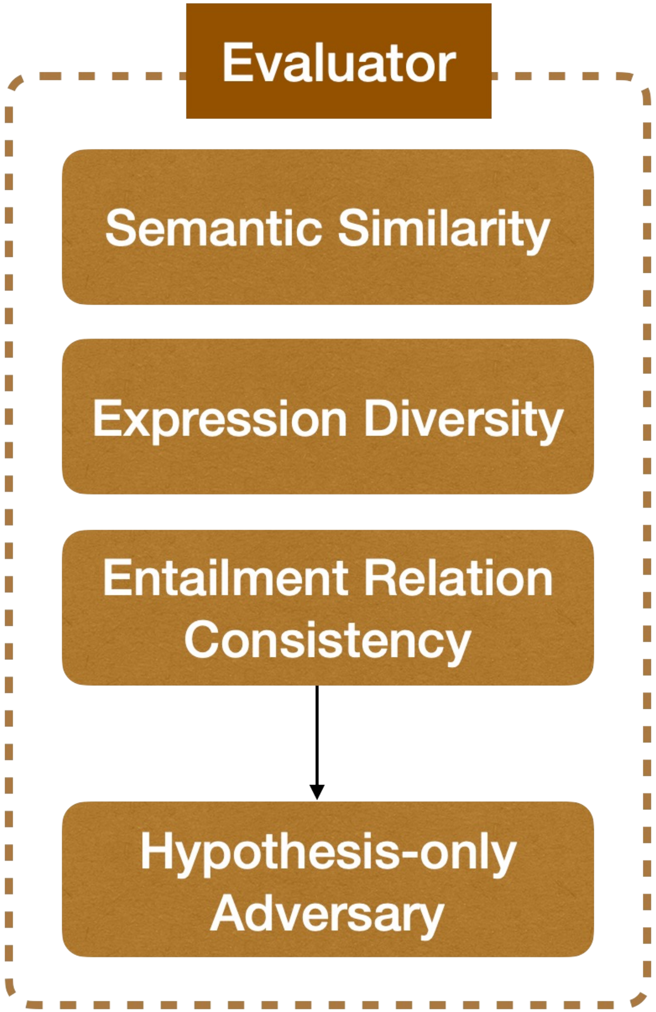
ERAP: Entailment Relation Aware Paraphraser

Generator: Takes in a sentence and an entailment relation for conditioning the generation.

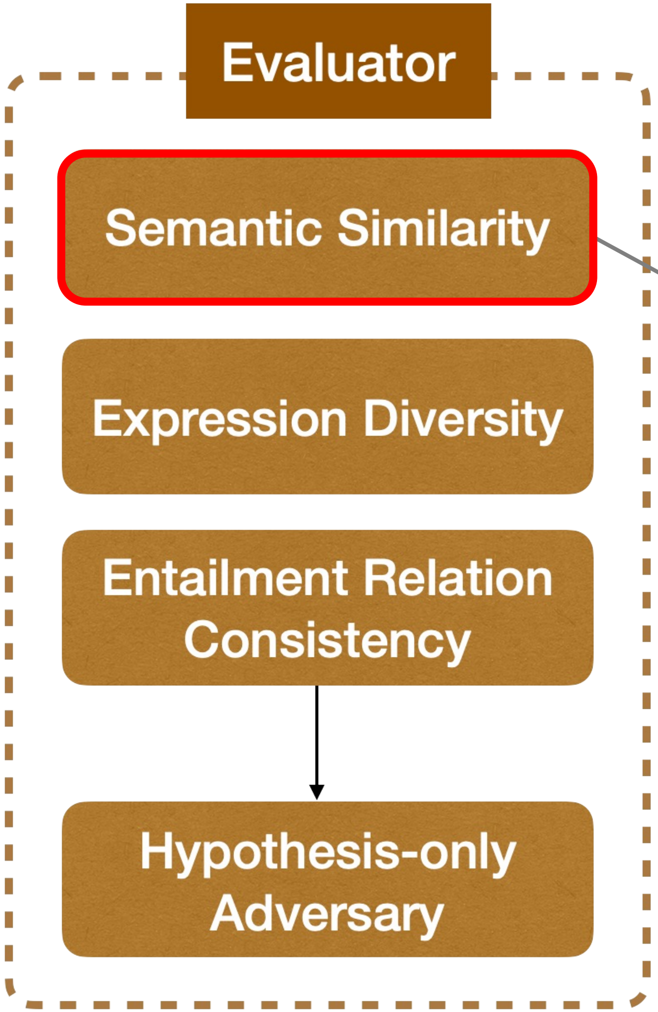


Evaluator: Takes in the generated paraphrase and scores it using different scorers to provide reward to the Generator.

Evaluator: Dive in

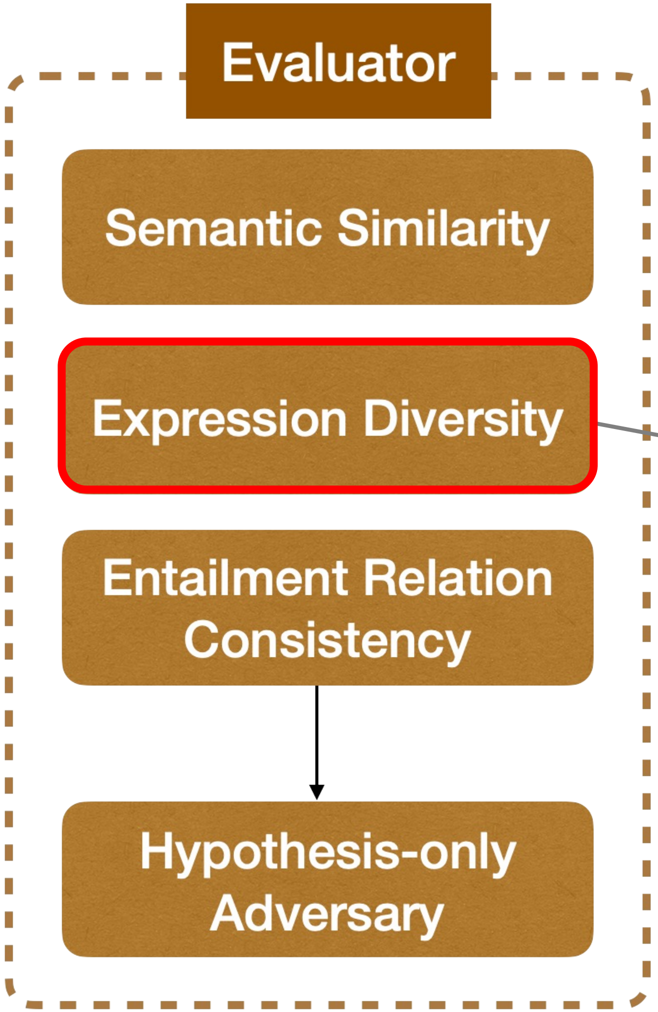


Evaluator: Dive in



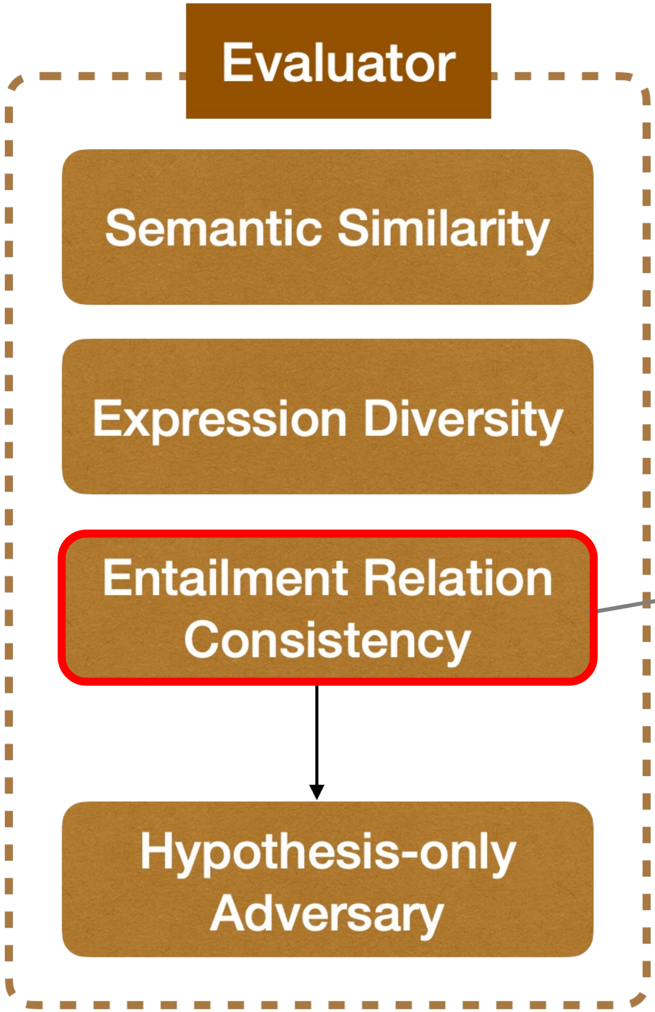
- Encourages generator to output paraphrases which are **closer in meaning** to the input
- MoverScore (Zhao et al., 2019) to measure closeness in meaning using word mover's distance between contextualized embeddings

Evaluator: Dive in



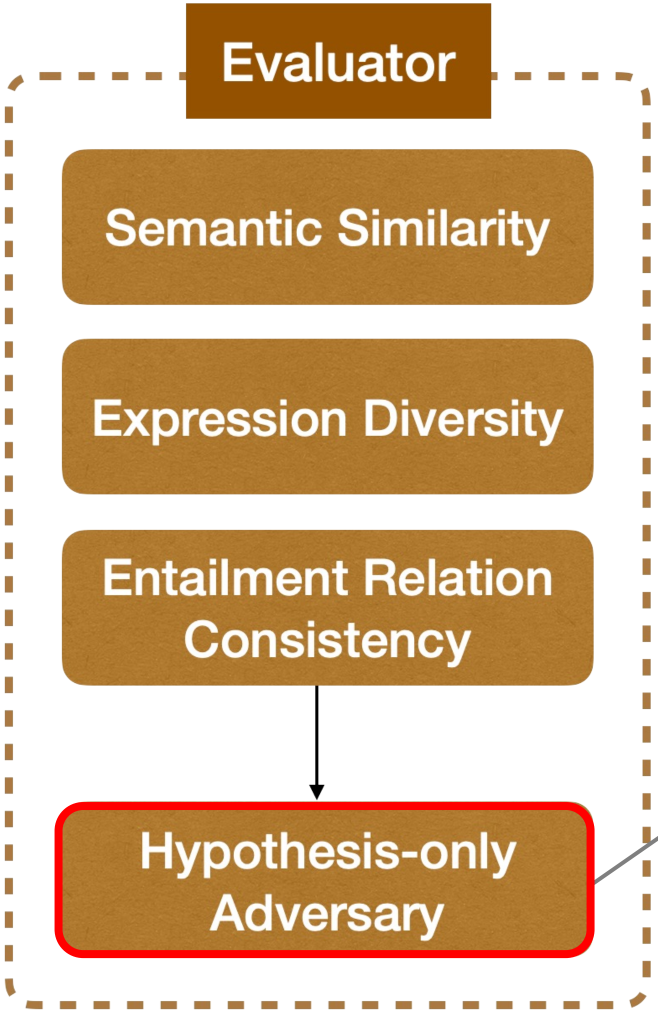
- Encourages generator to output paraphrases that **use different words** to express the input
- Inverse BLEU measures the diversity; 1 – BLEU(Y,X)

Evaluator: Dive in



- Encourages generator to output paraphrases **conforming to desired relation R**; O(X,Y) matches R
- Likelihood of input relation from Entailment Oracle measures the consistency

Evaluator: Dive in



- Generator may use certain strategies to get high consistency scores; insertion of tokens
- **Penalizes** paraphrases **if the desired relation can be predicted from the paraphrase alone**
- Trained adversarially alternating with the Generator

Entailment Relation Oracle

- NLI is a standard NLU task of determining whether a hypothesis *h* is true (entailment E), false (contradiction C), or undetermined (neutral N) given a premise *p*
- Train a natural language inference model from existing datasets (SNLI, MNLI, SICK, and HANS) to predict E, C, N
- Use NLI forwards (X,Y) and backwards (Y,X) to determine entailment relation label as follows:

$$O(X,Y) = \begin{cases} \equiv & \text{if } o(l|\langle X,Y\rangle) = E \;\&\; o(l|\langle Y,X\rangle) = E \\ \sqsubset & \text{if } o(l|\langle X,Y\rangle) = E \;\&\; o(l|\langle Y,X\rangle) = N \\ \sqsupset & \text{if } o(l|\langle X,Y\rangle) = N \;\&\; o(l|\langle Y,X\rangle) = E \\ \text{C} & \text{if } o(l|\langle X,Y\rangle) = C \;\&\; o(l|\langle Y,X\rangle) = C \\ \text{N} & \text{if } o(l|\langle X,Y\rangle) = N \;\&\; o(l|\langle Y,X\rangle) = N \\ \text{Invalid} & \text{,otherwise} \end{cases}$$

Evaluation

- Intrinsic
  - To assess the quality of generated paraphrases (Y)
  - To assess if the generated paraphrases (Y) conform to the desired relation (R)
- Extrinsic
  - To show benefits of entailment relation aware generation over unaware counterparts
    - Paraphrastic data augmentation: Augment training data for NLI downstream task
    - Assessing susceptibility to augmentation artifacts

Intrinsic: Comparison Models

- Supervised training on SICK in an entailment-aware (**Seq2seq-A**) and unaware (**Seq2seq-U**) supervised settings. (gold labels)
- Pre-training on ParaBank (Hu et al., 2019) dataset in entailment-aware (**Pre-trained-A**) and unaware (**Pre-trained-U**) settings. (weakly-supervised labels)
- Pre-training followed by supervised fine-tuning in entailment-aware (**Fine-tuned-A**) and unaware (**Fine-tuned-U**) settings.
- Use nucleus sampling (Holtzman et al., 2019) to generate k(=20) outputs from above unaware models and re-rank based on combined score from evaluator (**Re-rank-s2s-U; Re-rank-FT-U**)

Intrinsic: Automatic Evaluation

Model	$\mathscr{R}$-Test	BLEU↑	Diversity↑	*i*BLEU↑	$\mathscr{R}$- Consistency↑
Pre-trained-U	✗	14.92	76.73	7.53	–
Pre-trained-A	✓	17.20	74.25	**8.75**	65.53
Seq2seq-U	✗	30.93	59.88	17.62	–
Seq2seq-A	✓	31.44	63.90	**18.77**	38.42
Re-rank-s2s-U	✓	30.06	64.51	17.26	51.86
Re-rank-FT-U		41.44	53.67	**23.96**	**66.85**
ERAP-U[⋆]	✓	19.37	69.70	9.43	66.89
ERAP-A		28.20	59.35	**14.43**	**68.61**
Fine-tuned-U	✗	41.62	51.42	23.79	–
Fine-tuned-A	✓	45.21	51.60	**26.73**[*]	**70.24**[*]
Copy-input	–	51.42	0.00	21.14	45.98

lower-bound

upper-bound

***R*-Test**: If paraphrase is generated with relation control

-**U** without supervision/weak-supervision for entailment labels

-A with supervision/weak-supervision for entailment labels

Intrinsic: Human Evaluation

Model	**$\mathscr{R}$-Test**	**Similarity↑**	**Diversity↑**	**Grammar↑**	**$\mathscr{R}$-Consistency↑**
Pre-trained-U	✗	4.60	2.62	4.73	–
Pre-trained-A	✓	$\boxed{4.67}$	2.60	4.67	48.00
Re-rank-s2s-U	✓	2.72	$\boxed{3.15}$	3.46	24.00
Re-rank-FT-U		3.05	2.89	4.27	28.00
ERAP-U	✓	3.98	2.85	4.10	40.00
ERAP-A		3.95	2.68	4.42	$\boxed{64.00}$
Fine-tuned-U	✗	3.87	3.10	$\boxed{4.83}$	
Fine-tuned-A	✓	3.80	3.04	4.68	48.00

- 3 Mturk annotators per sample per metric
- Semantic similarity (alpha=0.65), Diversity of expression, (alpha=0.55), and Grammaticality (alpha=0.72) on 5-point Likert
- Entailment relation consistency (alpha=0.70): % of correct desired relation

Qualitative Outputs

1. Qualitative outputs for ablation study	
Input ⊏	a shirtless man is escorting a horse that is pulling a carriage along a road
Reference	a shirtless man is leading a horse that is pulling a carriage
Only Con	a shirtless man escorts it.
Con+Sim	a shirtless man is escorting a horse.
Con+Sim+Div	a shirtless man escorts a horse.
ERAP-A	a shirtless person is escorting a horse who is dragging a carriage.

Con: Consistency Scorer; **Sim**: Semantic similarity scorer; **Div**: Expression diversity scorer

Qualitative Outputs

2. Example of heuristic learned w/o Hypothesis-only Adversary	
Input ⊐	a man and a woman are walking through a wooded area
Reference	a man and a woman are walking together through the woods
-Adversary	a desert man and a woman walk through a wooded area
ERAP-A	a man and a woman are walking down a path through a wooded area
Input ⊐	four girls are doing backbends and playing outdoors
Reference	four kids are doing backbends in the park
-Adversary	four girls do backbends and play outside with mexico.
ERAP-A	four girls do backbends and play games outside.

Extrinsic Evaluation

- Paraphrastic Data Augmentation for NLI
  - Prior work (Hu et al., 2019a) has shown that paraphrastic data augmentation improves performance of NLI models, *but* assumes that entailment relations are preserved under paraphrasing which is not always the case

Extrinsic Evaluation

Given R(P,H), What is R(P`,H`)	H'=H	≡(H,H')	⊏(H,H')	⊐(H,H')
P`=P	E → E NE → NE	E → E NE → NE	E → E NE → U	E → U NE → U
≡(H,H')	E → E NE → NE	E → E NE → NE	E → E NE → U	E → U NE → U
⊏(H,H')	E → U NE → U	E → U NE → U	E → U NE → U	E → U NE → U
⊐(H,H')	E → E NE → NE	E → E NE → NE	E → E NE → U	E → U NE → U

E: Entailment   NE: Non-Entailment (Neutral or Contradiction)   U: Unknown

Extrinsic Evaluation

- Paraphrastic data augmentation for NLI
  - Prior work (Hu et al., 2019a) has shown that paraphrastic data augmentation improves performance of NLI models, *but* assumes that entailment relations are preserved under paraphrasing which is not always the case
  - *Hypothesis*: Entailment-aware augmentations result in reduced violation of labels and lead to better performance
- Assessing susceptibility to augmentation artifacts
  - Noisy training examples with incorrectly projected labels lead to *augmentation artifacts* in downstream tasks
  - *Hypothesis*: Models trained with entailment-aware augmentations are less susceptible to such *artifacts* than those trained with entailment-unaware augmentations. We create adversarial test set to investigate this.

Extrinsic Evaluation: Results (Accuracy)

Data	$\mathscr{R}$-Test	Original-Dev↑	Original-Test↑	Adversarial-Test↑
SICK NLI	-	95.56	$\boxed{93.78}$	$\boxed{\mathbf{83.02}}$
+FT-U(≡)	✗	95.15	93.68	69.72
+FT-A(≡)	✓	95.35	**94.62**	77.98
+FT-A(≡, ⊐)	✓	**95.76**	93.95	75.69
+ERAP-A(≡)	✓	95.15	94.58	78.44
+ERAP-A(≡, ⊐)	✓	95.15	93.86	69.72

***R*-Test**: If paraphrase is generated with relation control
Adversarial-Test: Incorrectly labeled paraphrastic augmentations in Test-set of SICK NLI

Takeaway

- Developed an RL-based model (ERAP) to generate paraphrases with controllable entailment relations ($\sqsubset$, $\sqsupset$, $\equiv$)
- ERAP uses NLI-trained oracle in lieu of labeling large paraphrase datasets with entailment labels.
- ERAP can be used for paraphrastic data augmentation while reducing augmentation artifacts.
- Entailment-aware paraphrasing provides control over the semantic nature of paraphrase; enhancing the applicability to downstream tasks

References

- Hu, J. E.; Rudinger, R.; Post, M.; and Van Durme. *ParaBank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural ma-chine translation*. In Proceedings of the AAAI, 2019.
- Wieting, J., & Gimpel, K. *ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. ACL, 2018*
- Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi. *The curious case of neural text degeneration, 2019.*
- Marelli, M.; Menini, S.; Baroni, M.; Bentivogli, L.;Bernardi, R.; Zamparelli, R. *A SICK cure for the evaluation of compositional distributional semantic models*. LREC, 2014.
- Hu, J. E.; Khayrallah, H.; Culkin, R.; Xia, P.; Chen, T.; Post,M.; and Van Durme, B. 2019a. *Improved lexically con-strained decoding for translation and monolingual rewriting.* In Proceedings of NAACL-HLT, 2019.
- Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and SteffenEger. *Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance.* In Proceedings of the EMNLP, 2019

Thanks!

Contact: sancheti@umd.edu