Entailment Relation Aware Paraphrase Generation

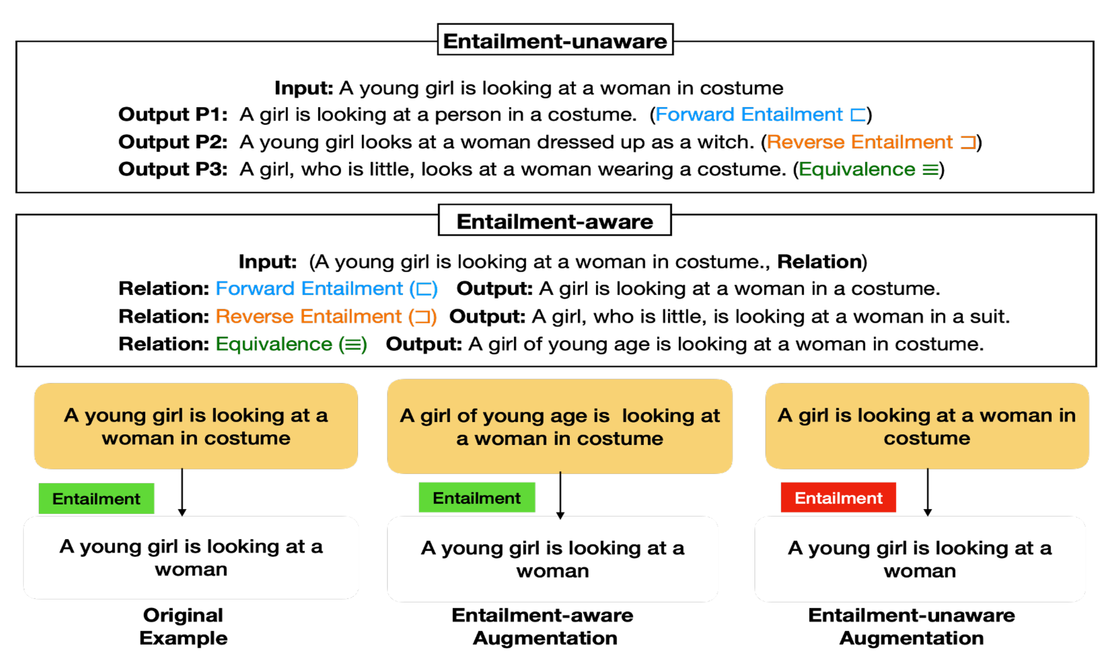
Abhilasha Sancheti[1,2], Racehl Rudinger[1], and Balaji Vasan Srinivasan[2]

[1]University of Maryland, College Park and [2]Adobe Research

Objectives

- We introduce a new paradigm of paraphrase generation with controllable entailment relations.
- We develop an RL-based paraphrasing system (**ERAP**) which can be trained using existing paraphrase and natural language inference (NLI) datasets.
- We build a NLI-trained oracle to obtain weak-supervision for entailment relation labels for existing paraphrase datasets.
- ERAP can be used for paraphrastic data augmentation while reducing augmentation artifacts.

Motivation



- Existing paraphrasing systems are unaware of the entailment relation between the generated paraphrases and the input.
- Such paraphrases when used to generate label preserving augmentation for downstream task such as textual entailment, might result in incorrectly labeled data.
- Explicit control over the entailment relation between the input and its paraphrase helps in reducing such incorrect data augmentations.
- $\equiv$ paraphrases useful in highly conservative and precise rewriting, $\sqsubset$ in summarization or simplification, and $\sqsupset$ in conversational AI.

Problem Definition

Given an input sentence X, and an entailment relation $\mathcal{R}$, generate a paraphrase Y such that the entailment relationship between X and Y is $\mathcal{R}$. We consider 3 relation controls.

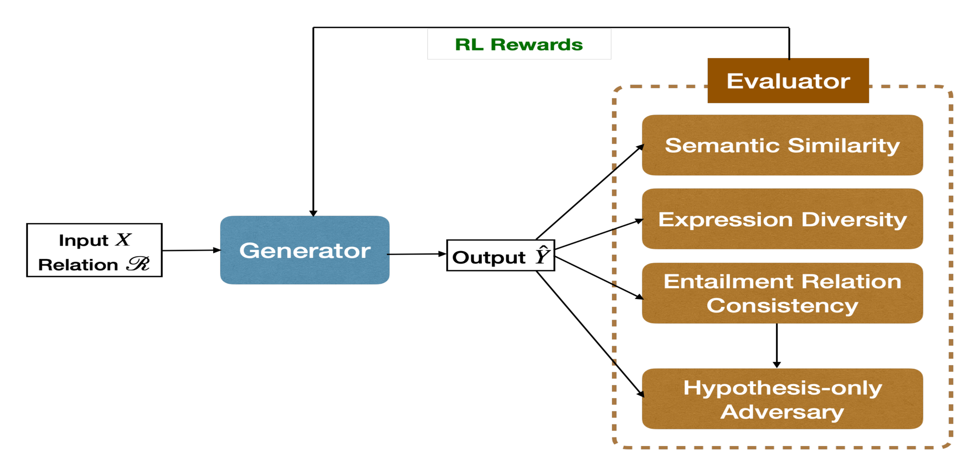
- Forward Entailment $X \sqsubset Y$:= if X is true, then Y is true.
- Reverse Entailment $X \sqsupset Y$:= if Y is true, then X is true.
- Equivalence $X \equiv Y$:= X is true if and only if Y is true.

Addressing Data Annotation Challenge

Need entailment relation labels for paraphrases to provide supervision. Three ways to address this challenge.

- **Recasting-SICK [1]:** To obtain gold entailment relation labels for meaning preserving sentence pairs.
- **Entailment Oracle:** NLI-trained Oracle to obtain weak supervision for entailment relation labels.
- **ERAP:** RL-based paraphraser trained using existing paraphrase and NLI datasets (SICK, SNLI, MNLI, HANS).

Entailment Relation Aware Paraphraser



- **Generator:** A seq2seq transformer pre-trained on ParaBank [2] or ParaNMT [3] to generate paraphrases $\hat{Y}$ given X and $\mathcal{R}$
- **Evaluator:** Consists of several scorers to score the generated paraphrases for quality and its consistency with the input relation.
  - **Semantic Similarity:** To measure closeness in meaning using MoverScore [4] which computes word-mover's distance between contextualized embeddings.
  - **Expression Diversity:** To measure use of different words, $1 - BLEU(Y, X)$
  - **Relation Consistency:** To encourage paraphrases which conform to input $\mathcal{R}$, $P_{oracle}(\mathcal{R}|(X, \hat{Y}))$
  - **Hypothesis-only Adversary:** To penalize if $\mathcal{R}$ can be predicted from $\hat{Y}$ alone. Trained alternating with the Generator.

Intrinsic Evaluation

- To evaluate quality of paraphrase and if it conforms to the desired $\mathcal{R}$

Model	$\mathcal{R}$-Test	BLEU↑	Diversity↑	iBLEU↑	$\mathcal{R}$- 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

Table: Automatic evaluation of paraphrases from ERAP against entailment-aware (A) and unaware (U) models. $\mathcal{R}$-Consistency is measured only for models conditioned ($\mathcal{R}$-Test) on $\mathcal{R}$ at test time. Shaded rows denote upper- and lower-bound models.

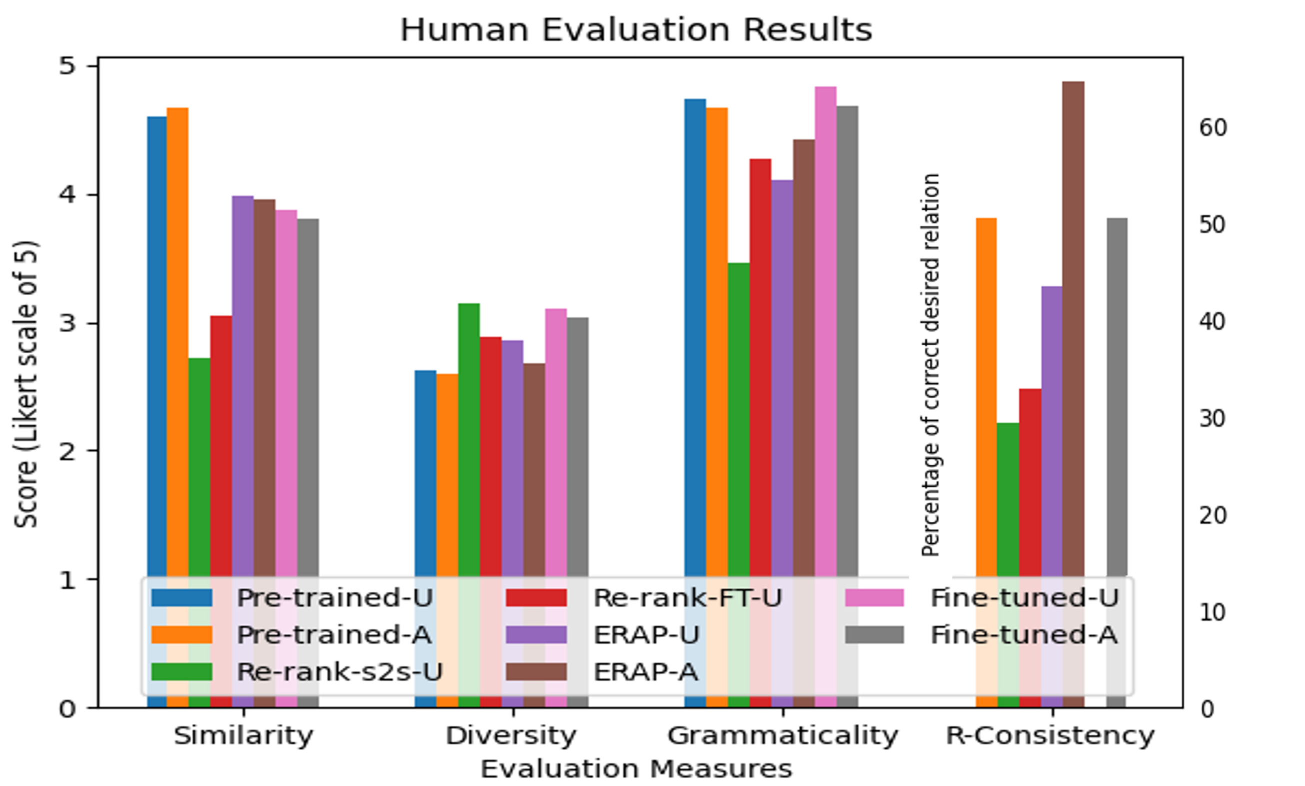


Figure: Mean across 3 annotators for Similarity (α=0.65), Diversity (α=0.55), Grammaticality (α=0.72) and % of correct relation for $\mathcal{R}$-Consistency (α=0.70).

Extrinsic Evaluation

- To show benefits of entailment-aware models over unaware models via paraphrastic data augmentation for textual entailment task.
- Naïvely assuming entailment label preservation under paraphrasing introduces incorrectly labeled (noisy) training examples leading to *augmentation artifacts* in trained models.

Data	$\mathcal{R}$-Test	Original-Dev↑	Original-Test↑	Adversarial-Test↑
SICK NLI	-	95.56	93.78	**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

Figure: Accuracy results: **FT**/**ERAP** refer to the Fine-tuned/proposed model used for generating augmentations. **U**/**A** denote entailment-unaware (aware) models. Improved performance of **-A** models over **U** while reducing *augmentation artifacts*.

References

[1] Marco Marelli, Stefano Menini, Marco Baroni, Bentivogli, Bernardi Luisa, Raffaella, and Roberto Zamparelli. A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, 2014.

[2] J Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of the AAAI*, 2019.

[3] John Wieting and Kevin Gimpel. Paranmt-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *ACL*, 2018.

[4] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the EMNLP*, 2019.