Supplementary for Paper: "UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation"

Jian Guan, Minlie Huang*
Department of Computer Science and Technology, Institute for Artificial Intelligence,
State Key Lab of Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
j-guan19@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn

A Negation Alteration Rules

We designed elaborate rules for negation alteration. The transformation rule from affirmative sentences to negative is shown in Table 1. In reverse, from negative sentences to affirmative, we removed the negation words ("not" or "n't") and altered the corresponding forms of the verbs.

Although there are other words which have negative meanings (e.g., "nobody"), they can be altered by another negative sampling technique: substitution with antonyms (e.g., replace "nobody" with "somebody"). Therefore, we did not process these words while performing negation alteration.

B Annotation Instruction

We show a screenshot of the annotation on AMT for a generated story given a leading context from ROCStroies in Figure 2. The annotation instruction for WritingPrompts is similar.

C Annotation Results

We averaged the scores of seven annotators as the final score for each story. Therefore, the annotation score ranges from 0 to 1 (i.e., $0, \frac{1}{7}, \frac{2}{7}, \cdots, 1$). The number distribution of stories with different scores is shown in Figure 1. Besides, we show 8 typical samples, one for each score in Table 2.

D Reconstruction performance

Besides the prediction objective, We also trained UNION with an auxiliary reconstruction task, which recovers the perturbation from a negative sample. During testing, we compute the Spearman correlation between human judgments and UNION's editing behavior. We measure the editing behavior by labeling 1 if UNION edits the input story, otherwise 0 if UNION just outputs the same story. The correlation is 0.1990 (p-value<0.01) on the whole test set of ROCStories, and is 0.3442/0.1652/0.1623/0.2943 on the evaluation set which contains repetitive/incoherent/conflicting/chaotic stories (the same setting with Table 7 of the main body, and each set is mixed with reasonable stories), respectively. Results show that it is easier to recognize the repetitive/chaotic stories, which agrees with the results in Table 7.

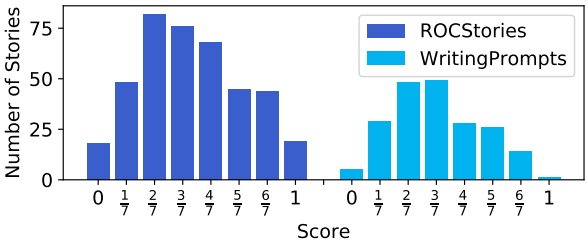


Figure 1: Number distribution of annotated stories with different human annotation scores. The total number for ROCStories/WritingPrompts is 400/200, respectively.

As for the editing output, although the key motivation of the reconstruction task is to provide more specific supervision signals for recognizing errors, UNION can generate meaningful editing results from unreasonable stories. We observe that UNION can correct lexical errors. For example, given the story "[FEMALE] worked real hard", UNION changed "real" to "really". However, since UNION adopted a non-autoregressive generative framework, it is difficult to generate a grammatical story if the input has sentence-level errors. But UNION can still accurately recognize the errors. For example, given the repetitive story "we had a great time. we had a great time.", it generated "we had a great time. we . .". We plan to improve the design by aligning the input and output tokens and then auto-tagging with editing operations during

*Corresponding author

training with the reconstruction task in the future.

E Case Study

We present several samples based on ROCStories and the corresponding judgments of different metrics in Table 3. We can see that it is difficult for baseline metrics to recognize the possible issues in stories, which rate the typical unreasonable stories (S2-S5) even higher than the reasonable one (S1). In comparison, UNION judges the quality of a story more accurately regardless of whether it is similar to the reference, suggesting that UNION can alleviate the one-to-many issue more effectively than referenced metrics (e.g., MoverScore). For instance, although S2 maintains a reluctantly reasonable plot through the story except for a repetitive sentence, annotators still give it zero because there is no such repetition error in human-written stories. And UNION successfully recognizes the issue thanks to the proposed negative sampling techniques which mimic the errors commonly observed in NLG models. Therefore, UNION is more reliable for evaluating open-ended story generation.

F Error Analysis

Although UNION outperforms the state-of-the-art metrics, it needs to be noted that the correlation with human judgments is still at a low level. As shown in Table 4, we present some typical cases where generated stories are misjudged by UNION. Firstly, although the proposed perturbation techniques have provided many lexical and syntactic variations, it is still hard to recognize some errors such as semantic repetition and emotionally conflicting (S6-S9). Secondly, we observed that UNION may not predict some reasonable stories (e.g., S10). This could be because some perturbated stories are still reasonable. For example, exchanging the order of two sentences without specific temporal relation (e.g., "he had to go through a lot of training" and "he took a first responder's course") does not break the story's coherence. Training with such noisy samples may make UNION misjudge some reasonable stories. Therefore, as future work, it is worth to explore more perturbation techniques for negative sample construction to reduce noise and cover more error types that UNION fails to recognize. Besides, it is necessary to introduce external knowledge to help judge the logic of stories.

References

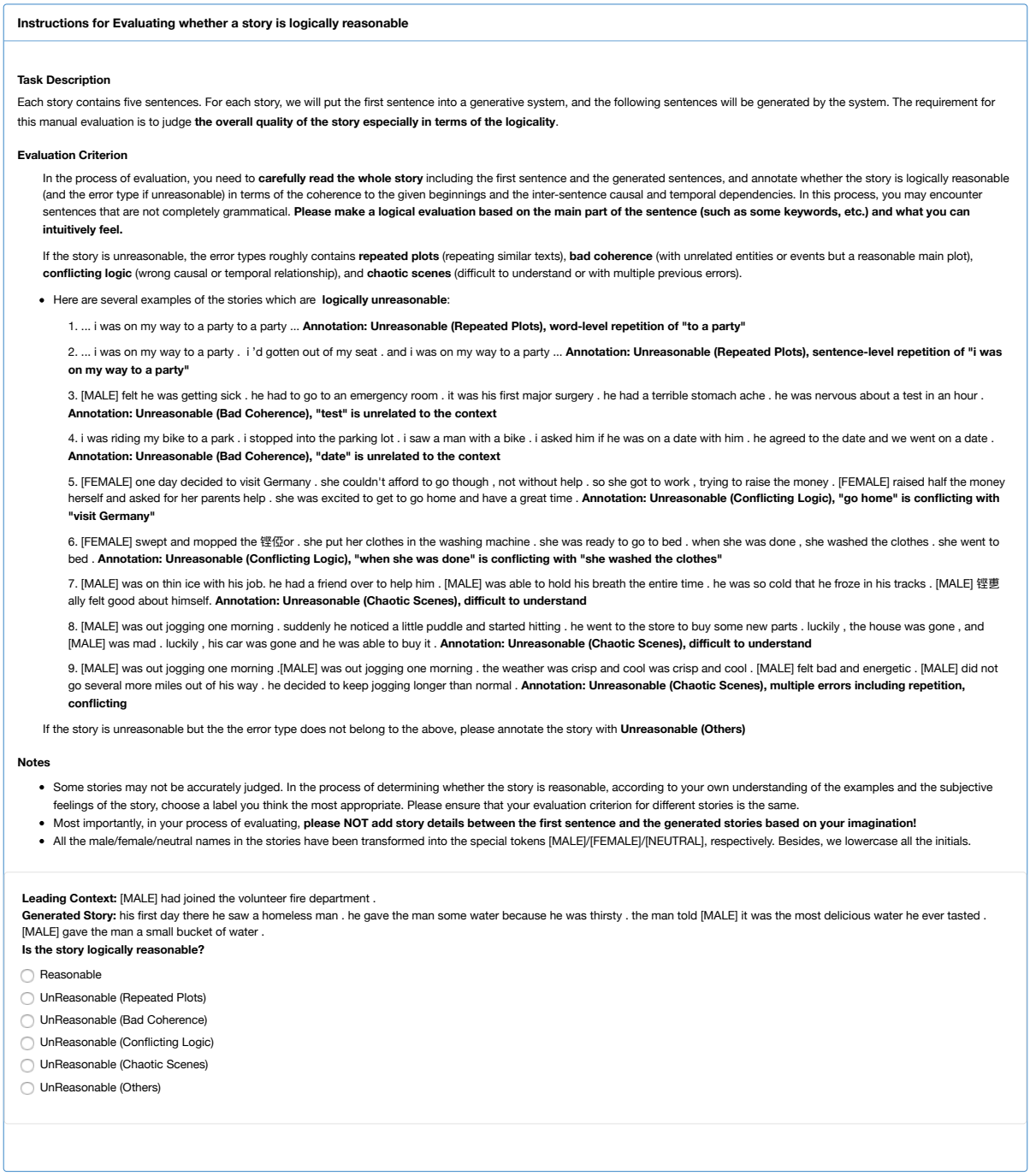
Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Verb Types	Verb Examples	Rules	Sentence Examples
be	am, was, been, ...	$\sim$ + not	Failure WAS NOT an option.
Modal verbs	would, will, shall, ...		I CAN NOT walk well.
Base	go	do not + v.	I DO NOT GO through the park.
3rd singular present	goes	does not + v.	He DOES NOT GO through the park.
Simple past	went	did not + v.	He DID NOT GO through the park.
Past participle	gone	not + $\sim$	His insurance rate had NOT GONE up.
Gerund	going		She ended up NOT GOING elsewhere.

Table 1: Transformation rules from affirmative sentences to negative by adding negation words for different types of verbs. $\sim$ stands for the current verb while **v.** for the base form of the verb. The CAPITALIZED words in sentence examples indicate the altered results. The negation word "not" can be randomly replaced with the short form "n't".

Scores	Samples	Repe	Cohe	Conf	Chao
0	**[MALE] went to work for his father's business.** He was very careful with his business. He *didn't get into trouble* for his mistakes. His father found out and *fired him*. He *was a bit sad* but *never did*.			✓	
$\frac{1}{7}$	**[NEUTRAL] and [MALE] had been dating for a little while.** *One day, [NEUTRAL] was convinced he could get a kiss. [NEUTRAL] decided to give her a kiss.* They agreed to *drink it* together. they had a good time at *the big bar*.		✓	✓	✓
$\frac{2}{7}$	**[FEMALE] noticed a bird's nest *by her bedroom window.*** She decided to *climb* the tree. She *climbed on the ladder* and *climbed* up to the window. She *climbed down the ladder* and *saw her step head*. She reached into *her pocket* and grabbed the bird's back.	✓	✓	✓	✓
$\frac{3}{7}$	**[MALE]'s narcissist girlfriend only cared about what he had to offer her.** *He was a successful businessman* who couldn't help but feed her desire. He did his best to show her that he was the real deal. She eventually left him because *he was a failure*. Although she left him, she never found someone else to love.			✓	
$\frac{4}{7}$	**[MALE] rented *an old apartment.*** He was very bored. He *was watching* a movie *while he watched* it. [MALE] asked the friend to watch it. [MALE] happily watched it.		✓	✓	✓
$\frac{5}{7}$	**One day [FEMALE] needed to *leave the airport.*** She was *waiting for her husband* to get out of work. He had a bad day at work. He asked her to *meet him at the airport.* [FEMALE] met her husband and they got in a taxi .			✓	
$\frac{6}{7}$	**[FEMALE] saw a *smoothie* at the store.** She saw a *chocolate cone*. she decided to buy it. She went and bought it. The *chocolate ice cream* was delicious.		✓		
1	**[MALE] had joined the volunteer fire department.** His first day there he saw a homeless man. He gave the man some water because he was thirsty. The man told [MALE] it was the most delicious water he ever tasted. [MALE] gave the man a small bucket of water.				

Table 2: Story samples for different human annotation scores and the annotated error types, including **Repe**ated plots, poor **Cohe**rence, **Conf**licting, and **Chao**tic scenes. **Bold** sentences are the given leading context. *Italic* words denote the improper entities or events.



Instructions for Evaluating whether a story is logically reasonable

Task Description

Each story contains five sentences. For each story, we will put the first sentence into a generative system, and the following sentences will be generated by the system. The requirement for this manual evaluation is to judge **the overall quality of the story especially in terms of the logicality**.

Evaluation Criterion

In the process of evaluation, you need to **carefully read the whole story** including the first sentence and the generated sentences, and annotate whether the story is logically reasonable (and the error type if unreasonable) in terms of the coherence to the given beginnings and the inter-sentence causal and temporal dependencies. In this process, you may encounter sentences that are not completely grammatical. **Please make a logical evaluation based on the main part of the sentence (such as some keywords, etc.) and what you can intuitively feel.**

If the story is unreasonable, the error types roughly contains **repeated plots** (repeating similar texts), **bad coherence** (with unrelated entities or events but a reasonable main plot), **conflicting logic** (wrong causal or temporal relationship), and **chaotic scenes** (difficult to understand or with multiple previous errors).

- Here are several examples of the stories which are **logically unreasonable**:

  1. ... i was on my way to a party to a party ... **Annotation: Unreasonable (Repeated Plots), word-level repetition of "to a party"**

  2. ... i was on my way to a party . i 'd gotten out of my seat . and i was on my way to a party ... **Annotation: Unreasonable (Repeated Plots), sentence-level repetition of "i was on my way to a party"**

  3. [MALE] felt he was getting sick . he had to go to an emergency room . it was his first major surgery . he had a terrible stomach ache . he was nervous about a test in an hour . **Annotation: Unreasonable (Bad Coherence), "test" is unrelated to the context**

  4. i was riding my bike to a park . i stopped into the parking lot . i saw a man with a bike . i asked him if he was on a date with him . he agreed to the date and we went on a date . **Annotation: Unreasonable (Bad Coherence), "date" is unrelated to the context**

  5. [FEMALE] one day decided to visit Germany . she couldn't afford to go though , not without help . so she got to work , trying to raise the money . [FEMALE] raised half the money herself and asked for her parents help . she was excited to get to go home and have a great time . **Annotation: Unreasonable (Conflicting Logic), "go home" is conflicting with "visit Germany"**

  6. [FEMALE] swept and mopped the 锘倬or . she put her clothes in the washing machine . she was ready to go to bed . when she was done , she washed the clothes . she went to bed . **Annotation: Unreasonable (Conflicting Logic), "when she was done" is conflicting with "she washed the clothes"**

  7. [MALE] was on thin ice with his job. he had a friend over to help him . [MALE] was able to hold his breath the entire time . he was so cold that he froze in his tracks . [MALE] 锘崽ally felt good about himself. **Annotation: Unreasonable (Chaotic Scenes), difficult to understand**

  8. [MALE] was out jogging one morning . suddenly he noticed a little puddle and started hitting . he went to the store to buy some new parts . luckily , the house was gone , and [MALE] was mad . luckily , his car was gone and he was able to buy it . **Annotation: Unreasonable (Chaotic Scenes), difficult to understand**

  9. [MALE] was out jogging one morning .[MALE] was out jogging one morning . the weather was crisp and cool was crisp and cool . [MALE] felt bad and energetic . [MALE] did not go several more miles out of his way . he decided to keep jogging longer than normal . **Annotation: Unreasonable (Chaotic Scenes), multiple errors including repetition, conflicting**

If the story is unreasonable but the the error type does not belong to the above, please annotate the story with **Unreasonable (Others)**

Notes

- Some stories may not be accurately judged. In the process of determining whether the story is reasonable, according to your own understanding of the examples and the subjective feelings of the story, choose a label you think the most appropriate. Please ensure that your evaluation criterion for different stories is the same.
- Most importantly, in your process of evaluating, **please NOT add story details between the first sentence and the generated stories based on your imagination!**
- All the male/female/neutral names in the stories have been transformed into the special tokens [MALE]/[FEMALE]/[NEUTRAL], respectively. Besides, we lowercase all the initials.

Leading Context: [MALE] had joined the volunteer fire department .
Generated Story: his first day there he saw a homeless man . he gave the man some water because he was thirsty . the man told [MALE] it was the most delicious water he ever tasted . [MALE] gave the man a small bucket of water .
Is the story logically reasonable?

- ◯ Reasonable
- ◯ UnReasonable (Repeated Plots)
- ◯ UnReasonable (Bad Coherence)
- ◯ UnReasonable (Conflicting Logic)
- ◯ UnReasonable (Chaotic Scenes)
- ◯ UnReasonable (Others)

Figure 2: A screenshot of the annotation on AMT for manual evaluation.

ID	Leading Context	Reference	Generated Samples	H	M	B	U
S1	[MALE] had joined the volunteer fire department.	He had to go through a lot of training. He took a first responder's course. [MALE] was first to respond on a scene one time. He saved a man 's life.	His first day there he saw a homeless man. He gave the man some water because he was thirsty. The man told [MALE] it was the most delicious water he ever tasted. [MALE] gave the man a small bucket of water.	1.00	0.34	0.43	0.99
S2	We were looking for something fun to do on a Tuesday night.	We decided to see a new movie that was out. When we got there we found out the tickets were half price on Tuesdays. We decided Tuesdays will now be our standing date night. It is such a nice, fun, cheap night that we can look forward to.	My wife and I were so excited. We went to the mall. *We had a great time. We had a great time. (Repe)*	0.00	0.44	0.49	0.00
S3	[NEUTRAL] had a new baby brother.	The baby would cry all night. [NEUTRAL] wasn't able to sleep. [NEUTRAL] started to despise his brother. He asked his mom if he could move to his grandmother's.	He wanted to do something new. He was sad to see other kids and play his own. [NEUTRAL] had a great time. [NEUTRAL]'s dad decided to *go shopping. (Cohe)*	0.00	0.48	0.54	0.00
S4	[MALE] went to work for his father's business.	His father was the boss. [MALE] was lazy at work. Everyone was scared to tell his father. [MALE] continued to do a bad job.	He was very careful with his business. He *didn't get into trouble* for his mistakes. His father found out and *fired him*. He *was a bit sad* but *never did. (Conf)*	0.14	0.62	0.69	0.00
S5	[FEMALE]'s mom married [FEMALE]'s dad, and the two girls became stepsisters.	[FEMALE], 12, had grown up in a low-income single-parent household. But ani, 7, was wealthy and spoiled, so she was very bratty. At first she hated [FEMALE] and was always mean to her! But then, finally, the two girls began to become friends.	When their dad *left the house*, *he went to their room*. When he came back, he found them in the closet. He scolded them and grounded them for a year. The girls *weren't happy with their new stepmother. (Chao)*	0.00	0.45	0.52	0.00

Table 3: Judgments of different metrics for several typical generated samples. **H, M, B** and **U** stand for human ratings, MoverScore, BLEURT and the proposed metric UNION, respectively. *Italic* words denote the improper entities or events and the specified error type, including *Repe*ated plots, poor *Cohe*rence, *Conf*licting, and *Chao*tic scenes. For comparison, we normalize all the scores to the range of [0,1] following Tao et al. (2018).

ID	Leading Context	Reference	Generated Samples	H	U
S6	[MALE] wanted to impress his friends.	[MALE] knew how to juggle. [MALE] juggled for his friends. His friends enjoyed the show. [MALE] was happy to captivate his friends attention.	He decided to make a homemade pie for himself. He *invited all his friends* over for a party. Everyone showed up with a recipe. [MALE] was very happy and *invited all of his friends. (Rept)*	0.43	0.98
S7	[FEMALE] came home from school with a lot of homework.	She usually asks her older sister for help. [FEMALE]'s older sister wasn't home. [FEMALE] needed help with her math homework. She then decided to asked her mother for help with her math.	She didn't know what to do with it. She asked her mom to teach her. *[FEMALE] said she needed a dog. But [FEMALE]'s mom was a little afraid of him. (Cohe)*	0.29	1.00
S8	[FEMALE] was a very well off lawyer in santa fe.	She decided to help a charity in her local area pro bono . [FEMALE] went to the state capital on the charity's behalf. She argued for laws protecting nonprofits and the people they serve. [FEMALE] was very successful and felt proud of her accomplishments.	She never had a college degree and had *not studied well*. She *was very good at her job* but always felt like a fraud. A lot of people *were disappointed with her*. [FEMALE] was very upset and decided to major in philosophy. *(Conf)*	0.14	1.00
S9	[MALE] was very nervous.	The big day had finally come and it was time to pop the question. He held her hand, but she didn't know what was going to happen. [MALE] got down on one knee and asked her to marry him. With tears in her eyes she accepted and they embraced.	He was at the bar with his girlfriend when a man got into his car. He saw that he was going to be alone. The man was *shocked by the situation and asked if she was okay*. [MALE] *went to his office and got his wife's name. (Chao)*	0.29	1.00
S10	[FEMALE] was making coffee before going to work.	But she realized she wouldn't have enough time. So she left her house right away. But when she came back, she realized her stove was still on. So she bought a smoke alarm just in case it happens again.	She didn't put the lid on the pot. She accidentally used the pot on the stove. [FEMALE] burned herself. [FEMALE] is now more careful with her pot.	0.86	0.00

Table 4: Typical misjudgments by UNION. **H** and **U** stand for human ratings and UNION, respectively. *Italic* words denote the improper entities or events and the specified error type, including *Repe*ated plots, poor *Cohe*rence, *Conf*licting, and *Chao*tic scenes.