A Data

Scales of the corpora and its use are summarized in the Table 2.

A.1 American Literature Short Story (ALSS)

A.1.1 Circumventing the Copyright Issue

Regarding the copyright issue on Toronto Book Corpus, we decided to collect ALSS as a replacement. The number of sentences and articles (or short stories) is addressed in the paper. Since the short story library website (https://americanliterature.com/short-story-library, Figure 5) had no corresponding terms of use, we tried a lot to contact with the administrator (via the only provided contact point: twitter) but failed to get any response. But robots.txt did not restrict crawling, we considered it as a implicit consent to non-profit use.

A.1.2 Data Collection and Refinement

To prevent raising any traffic problem by brutal crawling, the crawling code was designed to scrape one article per second on the website. Using spaCy (Honnibal and Montani, 2017) each crawled article is spliced into sentences. Splicing all the articles into sentences with spaCy takes almost a day. This resulted in a total of 1.148 million sentences.

We connected the consecutive sentence one by one to make the sentence to sentence prediction form of the dataset avoiding the last sentence of a story connected to the start of the other story article. Then those dataset examples were split into 8:1:1 ratio to make train, dev, and test set. After that, we filtered outliers amongst the sentences that are too lengthy (more than 70 tokens) or short (less than 3 tokens) which results in 11.98% (train/val/test = 11.99%, 11.99%, 11.90%) excluded sentences.

A.2 HellaSWAG

For this corpus (or dataset) the only thing to be noted clear is that the test split of the corpus of our use is made out of original validation split. We alternatively allocated each example in the original validation split into new validation and test split of our usage because the original test split had no answer labels on the provided multiple-choice candidates.

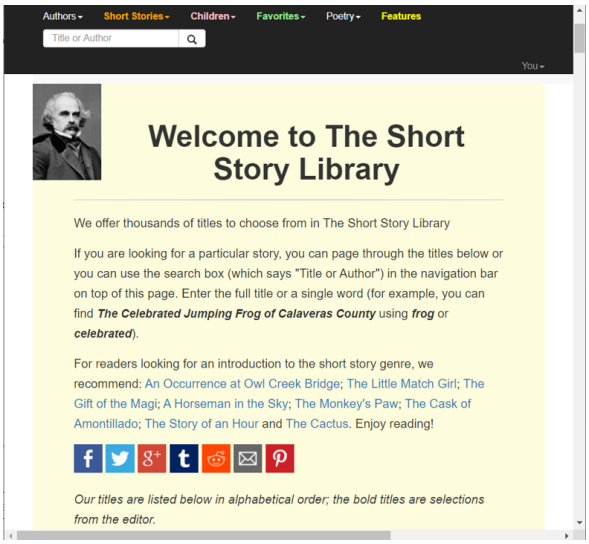

Figure 5: American Literature Short Stories web page, library section

A.3 Story Cloze Test

A.3.1 Dataset Version

Story Cloze Test has two versions: One is 2016 Spring and the other is 2017 Winter. Test and validation split of the corpus is only available on 2016 Spring version, we took this as our test and validation split of use. Train splits of both versions were merged to be used.

A.3.2 Fabricating Negative Examples on Training Split

For the training split, both versions had no incorrect endings (i.e., negative examples) for each 5-sentence stories, we had to fabricate a proper negative example for each as in validation and test split. To make this we took the correct endings of the other story as an incorrect ending for the current story. In this way, negative examples will also have the quality of human writing but with the wrong contents that make no sense. In the implementation, we set an offset of 8 to get a wrong ending (i.e., for story-1 we took the ending of story-9 to make a wrong ending to be used as a negative example).

A.3.3 Sanity Check for Fabricated Negative Examples

After fabrication of the negative example, we tested those toward human readers to see how many of the fabricated negatives confused the reader from original target story continuation. We tested toward 3 native speakers with 30 sampled binary choices from the resulted train split. Participants could distinguish a total of 96.7 % of the correct choices

(30/30, 29/30, 28/30) which is only less than 4 % of the fabricated data distorted the original answer label (in the original paper, cloze test performance by human testers was 100 %).

B Details of Human Evaluation

B.1 Job Qualification and Payroll

Survey participants are required to be capable of daily conversation in English and having AMT job approval rate over 92 percent. Each participant is paid with 5.00 USD per survey considering USA 2020 average minimum wage: 9.14 USD/hour.

B.2 Ethics Board

The platform we used for sentence assessment is AMT. We tried best not to collect any identifiable information of the individuals according to the terms of use. We also followed the IRB guidelines of S. Korea, which exempts approval of the experiments in our case.

B.3 Statistics

(Table 4, 5, 3 and Figure 6)
We provide details of the human evaluation reported in the paper. An actual screen of the survey the annotators saw (Figure 6), example survey question and choices (Table 3), summary statistics (Table 4), and break-down statistics according to corpora: HellaSWAG, Story Cloze Test (Table 5).

C Training Configurations

C.1 Training and Model Selection

(Table 6)
We choose Transformer (Vaswani et al., 2017) architecture for encoder-decoder LM. The dimension for the hidden and embedding vectors, and the feed-forward layer are reduced to 512 and 1024 respectively considering the small scale of our corpora.

For hyperparameter optimization, we focused on exploring learning rate and λ (for WR training) under several settings of warm-up steps and embedding mappings ($g(\cdot)$): linear, Hadamard product, and identity. We use Adam optimizer (Zaheer et al., 2018) with warm-up stage and a linear decay scheduler.

For model selection, we pre-selected two model checkpoints according to BLEU-1 score and validation loss, and chose one by reading generated examples. Pre-training of the encoder-decoder model performed on the ALSS dataset and fine-tuning the pre-trained model to the adapting corpora took less than 1 day and 8 hours, respectively, on one Titan Xp. All the corpora were tokenized by the punkt tokenizer provided by NLTK v 3.5 (Loper and Bird, 2002). Configuration details are in Table 6.

C.2 Experiment Configuration

(Table 6, Figure 7)
We mostly focused on tuning warm-up steps and learning rate. The max epoch of the fine-tuning phase of each model was unified to be 5 epochs which considered enough for fitting the model to the corpus. A minibatch size for each mode was adjusted according to the memory requirements of each training method where the WR training requires approximately two-times larger memory.

Additionally, we attach screenshots (Figure 7) of the WandB used for hyperparameter tuning which lead us to the configuration we noted here. For acquiring those values we ran around 500 experiment runs. The reason for using small that fits with one Titan Xp comes from here. The amount of the experiment runs were possible because WandB provides Sweep functionality which is search automation of the hyperparameter on several machines with a summary (https://docs.wandb.com/sweeps). We used Bayesian search mode provided rather than a grid search to reduce the number of runs.

D Generation Examples

(Table 7, 8, 9)
Other than introduced in the paper, we provide additional examples for more insight. We provide generation examples of WR winning over UL in preference rate in Table 7, Table 8, and also visit the opposite case in Table 9. One can see the n-gram score does not accord with the preference or quality well. The sum each preferred ratio in each example is not 100.0 % since the polls for *BOTH GOOD* or *NEITHER* are not appearing here.

E Ablation Studies

We provide ablation study that was decisive for choice of distance metric (d_{cos} vs. d_{euclid}), and representation mappings ($g(\cdot)$: linear, $\mathrm{Hadamard}$, $\mathrm{identity}$). Note that we could not see any reduction in validation loss ($TP(\cdot)$) for identity mapping (directly applying triplet loss to [CLS] hiddens); Use of learnable embedding mapping ($g(\cdot)$) helped WR training.

Also we provide the additional evidence for choosing repetitive 4-gram (denoted as "-rep.4" in Table 1) as a unlikely candidates, which is crucial for UL approach.

E.1 Choice of Embedding Mapping, $g(\cdot)$

(Table 10)

E.2 UL training

(Table 11)
As WR training only deals with sentence level fine-tuning, we used sequence level UL to unify the comparison environment which also reported to be effective on improving generation quality. On Table 11 we also report performance of the UL trained models according to unlikely token settings. As one can make own criterion for setting unlikelihood candidate n-grams for UL training, we tested random, negative unigram and repetitive 4-gram that scored best in the original paper.

F Additional Results

F.1 Automated Evaluation: BLEU and METEOR

(Table 12, Figure 8)
We used the same implementation of BLEU and METEOR from MSCOCO caption evaluation server (https://github.com/salaniz/pycocoevalcap). Table 12 contains BLEU-k and METEOR scores of the models toward two adaptation corpora. To note, "Single" baseline in the table is the same encoder-decoder model trained only on adaptation corpus (HellaSWAG and Story Cloze Test datasets) without pre-training on ALSS. Last row is pre-trained model performance on the ALSS corpus.

We pre-selected the models based on BLEU score and loss values but final decision was made with winnowing the generation samples. Note that BLEU scores are not normalized, which means the performance gaps between the models are closer than they look.

According to Novikova et al. (2017), n-gram metrics are not optimal for evaluating the performance of open-ended NLG tasks. To re-affirm, we also quantified how good the n-gram metrics are for predicting the sentence preference as shown in Figure 8.

F.2 Analysis on Learned Sentence Embeddings

F.2.1 Triplet Loss and Cosine Distance

(Figure , Table , and)

F.2.2 PCA Visualization

(Figure 10)
For extensive analysis, we provide PCA projections of the sentence representations learned by each model, for each adaptation corpus. Only HellaSWAG case shows distinguishable visual difference between generated continuation according to the training method[10]. WR-trained embeddings are closer to the embeddings of human-written sentences while UL or CE trained forming diverged lobe. We consider spurious projection for Story Cloze Test is caused by less informative negative examples which is shuffled continuation to other stories.

G Figures and Tables for Appendices

[10] We expected for more notable segregation of the embedding lobes between positive, generated, and negative continuation but couldn't see anything similar for other embedding methods such as t-SNE (Maaten and Hinton, 2008) and UMAP (McInnes et al., 2018).

Dataset	Original task	Use	number of data instance by splits (train/val/test)	number of negative examples per instance
ALSS	-	pre-training	808.34 k / 10.11 k / -	-
HellaSWAG	Next sentence prediction	text continuation	39.91 k / 5.02 k / 5.02 k	3
Story Cloze Test	Story completion	text continuation	98.16 k / 1.87 k / 1.87 k	2*/1 (train/val and test)

Table 2: Dataset description of ALSS, HellaSWAG, and Story Cloze Test used for pre-training and the text continuation task. Note that Story Cloze Test uses **fabricated** negative examples (*) in the training split.

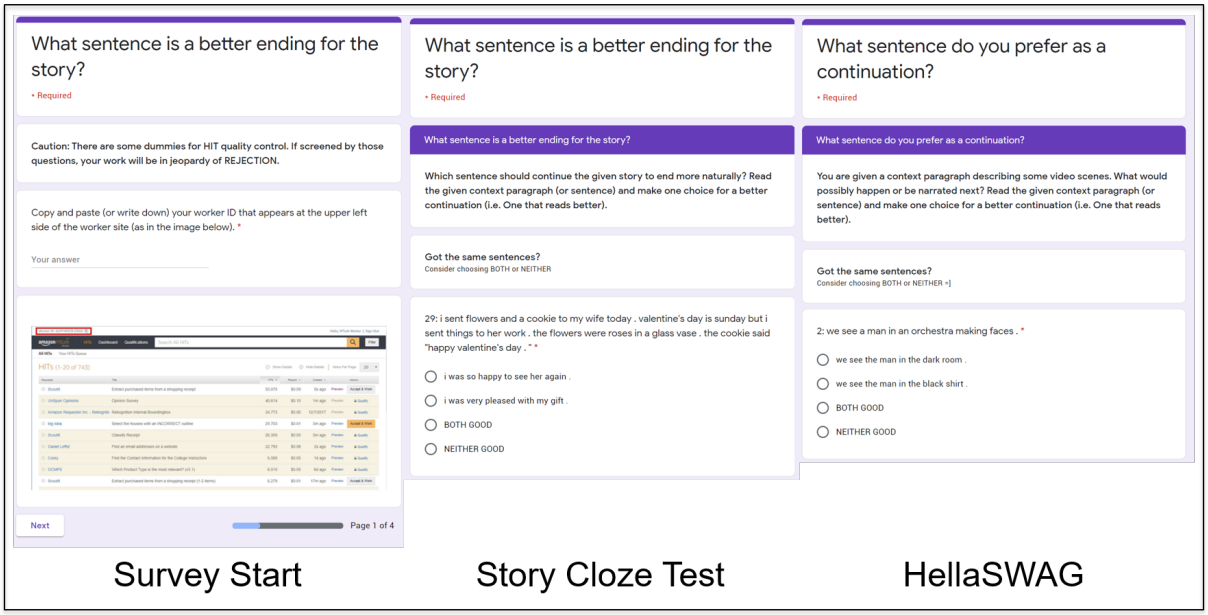

Figure 6: Survey screen that annotators watch. To verify each individual we used workerID on mTURK. We made a google form and linked with the AMT survey with verification code at the end of the survey.

Context Sentences	Choices
A man breaks into a house and begins to take things. he takes jewelry and games, some cash, and some food. when the family comes home they call the police. the police come and investigate and manage to track him.	the police officer is arrested.
	the police come and take the man to jail.
	BOTH GOOD
	NEITHER GOOD

Table 3: Question of the survey for evaluation. Annotators are asked: "What sentence do you prefer as a continuation?".

Total no. submission	156
Total no. valid submission	119
Total no. individuals of valid submission	107
Total no. questions (evaluated samples)	377
no. Questions per survey form (total 14 splits, except dummies)	26-28
Average no. workers per question	8.509
Average Inter-answerer agreement (%)	73.91%

Table 4: Summarized statistics of the human evaluation (survey). Inter-answerer agreement is preference share of the most preferred choice for each question.

	WR vs UL		WR vs CE	
	HellaSWAG	Story Cloze Test	HellaSWAG	Story Cloze Test
average number of valid workers per question	9.67	6.75	11.67	7
agreement in answerers	0.6937	0.7758	0.669	0.7909
no. questions	85	108	79	105
no. valid answerers	29.01	27	35.01	28

Table 5: Statistics break-down over adaptaion corpora. According to the corpus and training approaches (UL, CE), survey was performed on Amazon Mechanical TURK platform. There are 26-28 questions for survey form (i.e., For HellaSWAG corpus we distribute 3 versions of survey per each column, and for Story Cloze Test corpus 4 versions of survey forms were used).

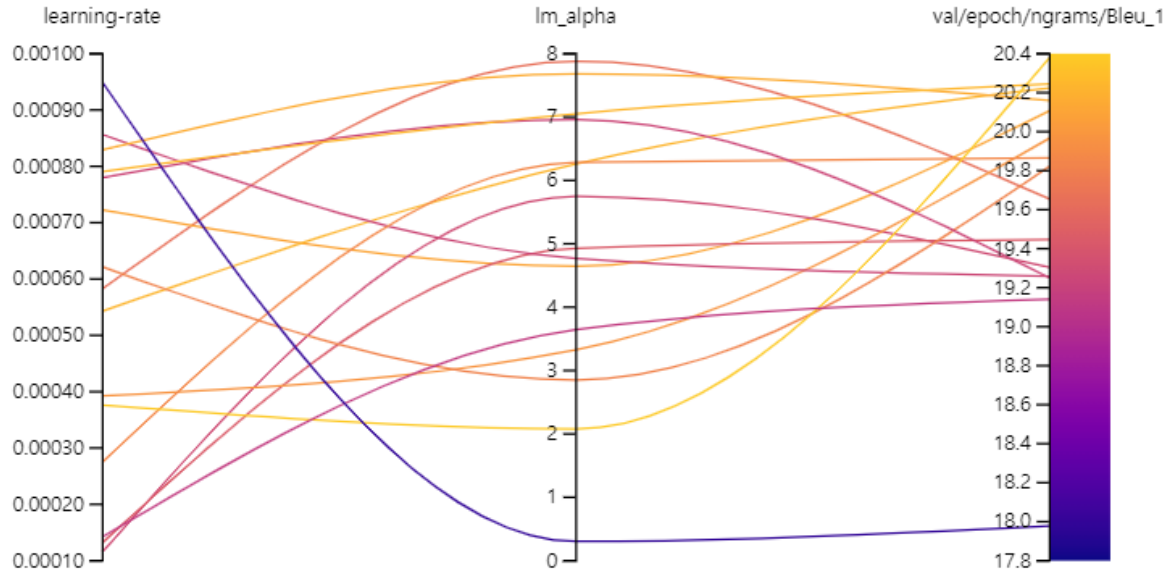

Figure 7: Sweep search of the WR training (distance: cosine, mapping: Hadamard product) on HellaSWAG corpus. λ and learning rate is being searched within Bayesian sweep based on resultant validation BLEU-1 score.

Corpus	Training	Learning rate	warm-up	Batch size	Max/Stopped ep	etc.
HellaSWAG	WR (cos-linear)	8.292e-4	2000	16	5/5	λ =7.683
	UL (repeat 4)	4.792e-4	2000	32	5/5	sequence tune rate: 0.5
	CE	5.376e-4	2000	16	5/5	
	Single	1.913e-4	11013	64	20 / 5	
Story Cloze Test	WR (cos-elem)	4.782e-4	2000	16	5/5	λ =6.957
	UL (repeat 4)	3.906e-4	2000	32	5/5	sequence tune rate: 0.5
	CE	5.517e-4	2000	16	5/5	
	Single	9.946e-4	10034	64	20 / 8	
ALSS	Pre-trained	4.751e-4	32956	32	10 / 9	

Table 6: Training conditions in detail. (cos-linear/elem) are notating metric space used and following embedding mapping ($g(\cdot)$) used for WR loss, and (repeat 4) means repeating 4-gram is considered as unlikelihood candidate for UL setting. Common factors such as hidden and embedding dimension (=512), ($\beta_1, \beta 2$)=(0.9, 0.99) for Adam optimizer, number of attention blocks (=6) and vocabulary frequency threshold (=4) are obmitted here. For *pre-trained* and Single models, both learning rate and warm-up steps were explored. For the others, warm-up steps of 2000, 4000, 8000, and 16000 were tested and fixed to 2000 and then searched for optimal learning rate and λ. Unlikelihood training here is sequence level only and followed the original tuning rate of 0.5.

Corpus	Context Sentence(s)	WR (proposed)	UL-rep.4
HellaSWAG	96: people are lined up to race bicycles . they race bicycles around the track .	(B1: 42.9, M: 20.1, **87.5 %**) they are going down the track .	(B1: 22.2 , M: 5.8 , 0.0 %) they are going to jump in the race .
HellaSWAG	22: a group of people are sitting in their donut tubes up on a snowy mountain .	(B1: 44.4 , M: 19.7 , **100.0 %**) they are then seen riding down the hill .	(B1: 44.4 , M: 14.3 , 0.0 %) they are then shown riding down the river .
HellaSWAG	77: how to plan a memorable marriage proposal . friends and family who have known your partner for a long time will likely have an idea if they have a dream proposal , and what it might include . for instance , your partner's siblings may recall your partner talking about how wonderful it would be to receive a marriage proposal on a beach , at sunset .	(B1: 40.9 , M: 7.5 , **100.0 %**) if you are not sure what your partner is doing , you may want to make a list of your own feelings.	(B1: 33.3 , M: 8.2, 0.0 %) if you are not sure what your partner is doing , you may want to be able to tell them that you are not
HellaSWAG	14: a woman is shown preparing to lift a barbell into the air . she lifts it into the air and above her head . she drops the barbell to the ground .	(B1: 33.3 , M: 5.8 , **69.2 %**) she then demonstrates how to do the exercise .	(B1: 30 , M: 5.7 , 7.7 %) she then begins to jump rope and jump rope .
HellaSWAG	91: a dance floor is shown covered in ballerinas . a woman talks to the camera in an interview .	(B1: 16.7, M: 5.8 , **100 %**) the woman is holding a pose and talking to the camera	(B1: 20 , M: 3.3 , 0 %) the woman is shown playing a guitar and singing .
HellaSWAG	7: the players have a long volley until the play in the foreground misses and the birdie lands at her feet .	(B1: 45.5 , M: 5.7 , **15.4** %) the ball is thrown back and forth in the air	(B1: 50 , M: 4.7 , 0 %) the ball is thrown back and forth .
HellaSWAG	16: a design is drawn on the pumpkin and then cut out . the inside of the pumpkin gets sprayed with mold remover and patted dry .	(B1: 21.4, M:4.2, **69.2 %**) the person then cuts the pumpkin into sections and cuts it into pieces .	(B1:38.5, M:7.4, 7.7 %) the pumpkin is then shown again and the pumpkin is shown again .
HellaSWAG	90: the adult mows the lawn with his child . the man moves toys to the side of the yard to clear the area .	(B1: 28.6 , M: 14.9 , **100 %**) the man continues to play the piano while the camera captures him from several angles .	(B1: 18.8 , M: 9.3 , 0 %) the ball is thrown back and forth .
HellaSWAG	78: a person is seen lying on the ground with another person running from behind .	(B1: 35.7 , M: 14.4 , **37.5 %**) the person continues moving around and ends by holding a stick .	(B1: 41.7 , M: 14.6 , 25 %) the woman is shown playing a guitar and singing .
StoryClozeTest	10: ella's dad had to go on a business trip to munich . ella begged her dad to bring her with him . she had never been to germany and she wanted to see it ! her dad agreed and brought ella with him on the trip .	(B1: 10.0, M: 1.7, **90.0 %**) ella was so happy to have a great trip !	(B1: 10.0, M: 8.9, 0.0 %) ella was so happy to have a new job !
StoryClozeTest	57: my parents split up . my dad went out of state to work . he met a new girl . they fell in love .	(B1: 60 , M: 12.6 , **85.7 %**) they are now married .	(B1: 20 , M: 1.8 , 0 %) she was very happy .
StoryClozeTest	46: bryan took his girlfriend to the movies . they went to a comedy . he bought popcorn and drinks . they laughed a lot .	(B1: 50 , M: 6.3 , **88.9 %**) they had a great time .	(B1: 12.5 , M: 6.7 , 0 %) he was happy to have a movie .

Table 7: Generation example comparison (1) between WR and UL training. BLEU-1 (B1), METEOR (M), and preference ratio (%).

Corpus	Context Sentence(s)	WR (proposed)	UL-rep.4
StoryClozeTest	49: sam has a toothache . he makes a dentist appointment . sam is very nervous . sam's dentist is very nice and gentle .	(B1: 55.6, M: 3.7, **22.2 %**) sam is happy that he has a dentist .	(B1: 50.0, M: 36.9, 11.1 %) sam is happy that he has a new dentist .
StoryClozeTest	94: i found a cat outside one day . i took her to the boat with me . i was planning on taking her to the shelter . however , i got very attached to her .	(B1: 50.0, M: 2.0, **100.0 %**) i was glad to have a cat .	(B1: 30.8, M: 11.5, 0.0 %) i was so happy that i had to go to the hospital .
StoryClozeTest	7: every year we attend a poolside barbecue at my aunt's house . one year when my sister was very young she fell in the pool . as soon as my dad realized she was missing he ran for the pool . he dove in and brought my sister to the surface .	(B1: 50.0, M: 1.3, **80.0 %**) we were very happy to see her .	(B1: 12.5, M: 3.0, 20.0 %) she was very happy with the results .
StoryClozeTest	69: maddy was always used to seeing old man joe . she thought of him a sort of watchman . each morning when she passed him she 'd give him a dollar . sometimes , she 'd even bring him some coffee or a bite to eat .	(B1: 20.0, M: 8.0, **42.9 %**) she was so happy to have a new friend .	(B1: 20.0, M: 8.6, 14.3 %) he was so happy to have his own coffee .
StoryClozeTest	92: it was my first time ever playing football and i was so nervous . i had never been tackled before . i hoped it would not hurt . during the game , i got tackled and it did not hurt at all !	(B1: 10.0, M:5.9, **100%**) i was so relieved when i won the game !	(B1: 18.2, M: 8.1, 0 %) i was so scared that i fell in the ground .
StoryClozeTest	15: the bradley family decided to take a camping trip . they packed the car for the two hour trip . once they arrived it was almost dark and began to storm . after the rain had passed they set up the tent and started the fire .	(B1: 66.7 , M: 16.5 , **70 %**) they had a great time .	(B1: 12.5 , M: 2.2 , 0 %) they were exhausted when they got home .

Table 8: Generation example comparison (2) between WR training and UL training. BLEU-1 (B1), METEOR (M), and preference ratio (%).

Corpus	Context Sentence(s)	WR (proposed)	UL-rep.4
HellaSWAG	30: a person is shown vacuuming a large rug with a green vacuum cleaner .	(B1: 50.0, M: 12.1, 0.0 %) the person uses a brush to brush the carpet .	(B1: 44.4, M: 7.0, 84.6 %) the person then begins to mop the floor .
HellaSWAG	64: a large chandelier sparkles over a stage . people in black clothing appear , backs turned . a woman lifts a crown , then the man starts dancing .	(B1: 23.1, M: 11.3, 0.0 %) the man then begins to dance while the camera captures his movements .	(B1: 30.0, M: 9.2, 87.5 %) the man then begins to dance with the camera .
StoryClozeTest	6: mary liked jumping rope . she would do it all the time . it kept her in good shape . she always had energy .	(B1: 44.4, M: 15.0, 10.0 %) she was happy to be able to play .	(B1: 44.4, M: 15.1, 70.0 %) she was happy to have a new hobby .
StoryClozeTest	16: jon was getting out of shape . jon decided to lose weight . he began to eat and exercise better . jon kept it up and started to feel results .	(B1: 42.9, M: 13.2, 0.0 %) jon lost weight and lost weight .	(B1: 71.4, M: 21.7, 100.0 %) jon was able to lose weight .

Table 9: Generated examples that the proposed training method being worse in preference. BLEU-1 (B1), METEOR (M), and preference ratio (%).

Corpus	distance	mapping	Bleu_1	Bleu_2	Bleu_3	Bleu_4
HellaSWAG	cosine	**elem**	**21.18**	**8.47**	**3.36**	0.93
		linear	20.85	8.37	3.33	**0.95**
		none	21.52	**8.71**	**3.51**	**1.07**
	euclid	elem	20.66	8.22	3.26	0.88
		linear	20.41	8.00	3.07	0.83
		none	20.32	8.00	3.01	0.77
Story Cloze Test	cosine	elem	**25.85**	**9.49**	**3.90**	1.63
		linear	25.71	9.18	3.86	**1.70**
		none	25.65	**8.99**	**3.40**	**1.30**
	euclid	elem	25.38	9.05	3.62	1.62
		linear	25.20	8.98	3.50	1.42
		none	25.34	8.94	3.51	1.31

Table 10: Mapping and distance metric performance ablation: Overall BLEU score shows that cosine metric space is prefered. For mapping function $g(\cdot)$, Hadamard product and linear projections are competitive.

Corpus	Unlikely candidate (n-gram)	B-1	B-2	B-3	B-4
HellaSWAG	**repeat.4**	**20.88**	**8.49**	**3.50**	**1.16**
	negative.1	18.67	7.41	3.08	0.90
	random.1	17.27	6.50	2.47	0.65
Story Cloze Test	**repeat.4**	**25.27**	**9.00**	**3.51**	1.12
	negative.1	23.62	8.44	3.46	**1.39**
	random.1	20.85	7.62	3.23	1.36

Table 11: Performance report for sequence level UL training according to the unlikely candidate criteria; repeating 4-gram (repeat.4), negative 1-gram (negative.1), random 1-gram (random.1). As reported in the original paper, repeating 4-gram shows better performance.

Corpus	Training Loss	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR (v1.5)
HellaSWAG	WR (cos-elem)	21.182	8.467	3.357	0.927	10.086
	UL (repeat 4)	20.883	8.493	3.504	**1.164**	9.945
	CE	**21.599**	**8.800**	**3.538**	0.912	9.883
	Single	19.553	7.582	3.021	0.864	**10.175**
	Pre-trained	15.059	4.056	0.510	0.045	5.144
Story Cloze Test	WR (cos-linear)	**25.710**	**9.180**	**3.860**	1.700	5.287
	UL (repeat 4)	25.270	9.000	3.510	1.120	5.330
	CE	**25.710**	9.080	3.650	1.580	**5.602**
	Single	25.500	9.070	**3.860**	**1.900**	5.551
	Pre-trained	9.898	2.172	0.462	0.027	3.887
ALSS	Pre-trained	17.976	6.532	2.278	0.6425	-

Table 12: N-gram metric scores. For each adaptation corpus and mapping option $g(\cdot)$, the best-performing (BLEU-1) weights of each training objective are chosen. The BLEU and METEOR scores of WR, UL, and CE are quite close to each other. The difference between WR-trained model and the best model are at most 0.42 in BLEU-1 and 0.32 in METEOR. We report the BLEU score of the pre-trained model on the ALSS to verify the pre-training.

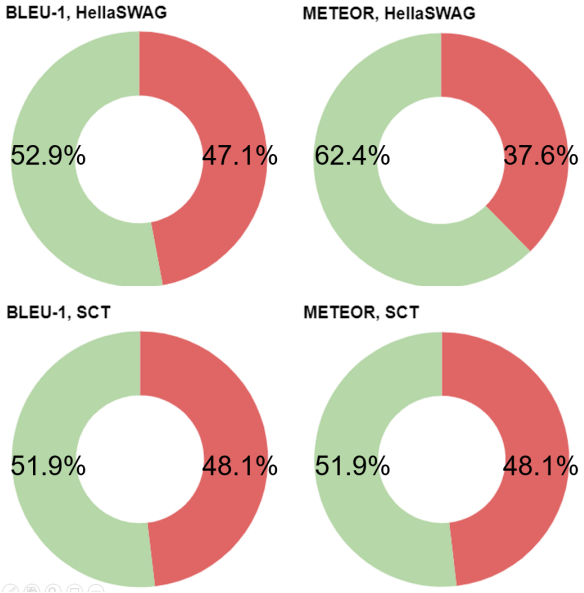

Figure 8: The accordance of the n-gram measures with human preference in UL vs WR (proposed) survey. The cases where the preference goes with n-gram score dominance (green) are slightly more than half except the METEOR-HellaSWAG case. Measured over generations appeared on WR vs. UL survey (85, 108 continuations and 822, 729 human preference judgements for HellaSWAG, and Story Cloze Test corpora respectively.)

	HellaSWAG (μ / σ^2)	Story Cloze Test (μ / σ^2)
WR	**0.885** / 0.017	1.007 / 0.074
UL	0.963 / 0.003	0.998 / >0.001
CE	0.954 / 0.003	0.998 / >0.001

Table 13: Triplet loss value (TP_{cos}) of each test split. The loss value ranges from 0 to 2. If the value is 0, the representation of the generated sentence perfectly aligns with the positive while opposite to the negative. More details can be found in Appendix F.2.1, Table 14.

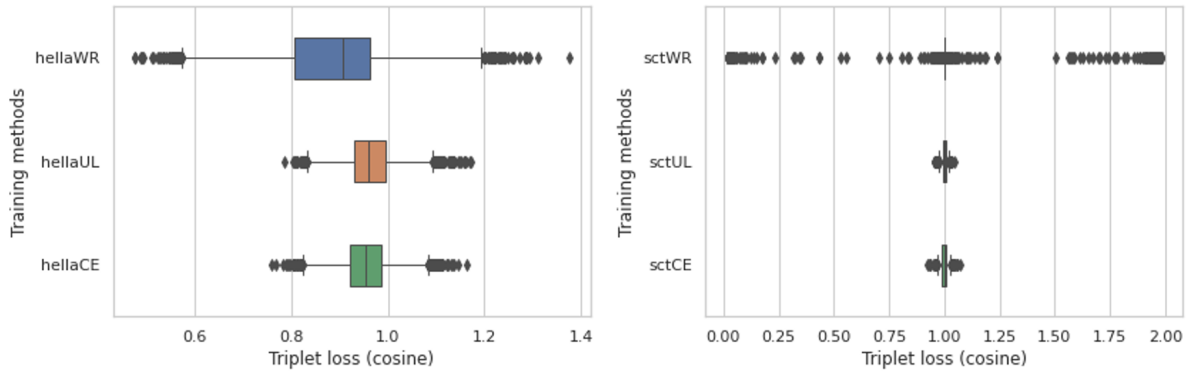

Figure 9: Whisker plot of the triplet loss ($\mathrm{TP}_{\cos}$). For HellaSWAG case, we see sentence representations of the WR generation samples resulting lower loss value. In Story Cloze Test case, however, median loss are not distinguishable while its variance is larger for WR training. We consider this is the effect of using noisy replacements of negative examples we prepared for training.

	HellaSWAG (μ / σ^2)		Story Cloze Test (μ / σ^2)	
	d^+_{cos}	d_cos^-	d^+_{cos}	d^-_{cos}
WR	0.166 / 0.010	0.281 / 0.011	0.116 / 0.091	0.109 / 0.087
UL	0.180 / 0.005	0.217 / 0.003	0.028 / >0.001	0.030 / >0.001
CE	0.178 / 0.005	0.224 / 0.003	0.039 / >0.001	0.041 / >0.001

Table 14: Cosine distances between generated sentences and positive (d^+_{cos}) and negative examples (d^-_{cos})

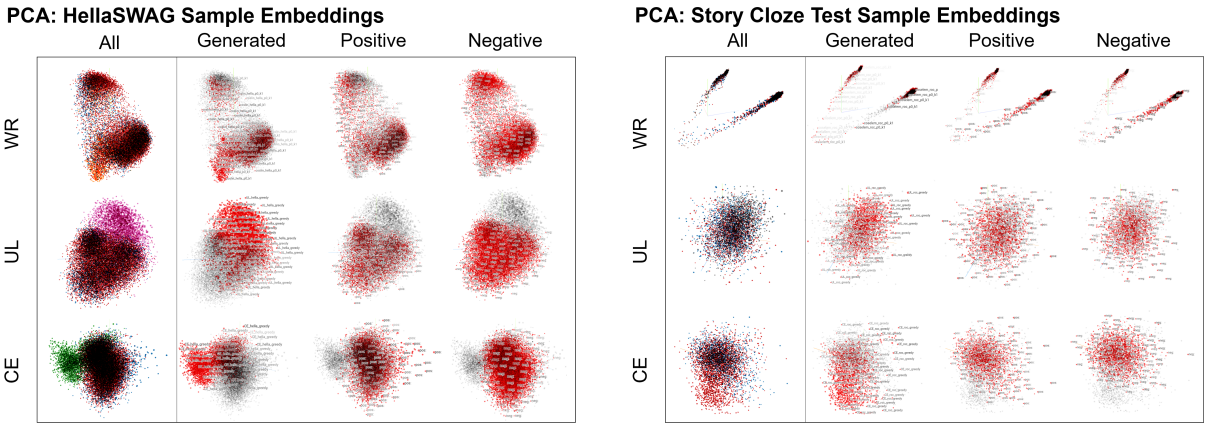

Figure 10: PCA projection of generated, positive and negative sentence (+context) embeddings for each training method. Upper plot is for HellaSWAG corpus and the other for the Story Cloze Test corpus. Each point in the figure is a projected sentence embedding. Except for the 1st column, red-colored points represent the following column groups; generated, positive, and negative. Rows represent training approaches.