Paper Reading

Jun Gao

2021-04-18

Overview

- A Study in Improving BLEU Reference Coverage with Diverse Automatic Paraphrasing
- Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining

Introduction

- Machine translation system outputs are usually evaluated against a single reference.
- This especially affects MT′ s dominant metric, BLEU...

Ref: This did not bother anybody .
MT_1: **This** didn ’t **bother anybody** .
MT_2: Nobody was bothered by **this** .

Introduction

- Unfortunately, multiple references are rarely available due to the **high cost and effort** of producing them.
- One way to inexpensively create them is with **automatic paraphrasing**.

Research Questions

- whether adding automatically paraphrased references can **provide the diversity** needed to better **cover the translation space**;
- whether this **increased coverage overlaps** with observed and valid MT outputs, in turn improving BLEU′ s correlation with human judgments.

Contributions

- explore two questions, testing on all into-English directions of the WMT19 metrics shared task at the system and segment level.
- compare two approaches:
  - 1) generate diverse references with the hope of covering as much of the valid translation space as possible. (diverse method)
  - 2) target the relevant areas of the translation space by generating paraphrases that contain n-grams selected from the system outputs. (non-diverse method)

Creating diverse paraphrases

- 1. Cluster target sentences by some property (e.g., semantic, syntactic representation);
- 2. Assign a code to each cluster and prefix each target sentence in the training data with its code (a pseudo-token)
- 3. Train an NMT-style paraphrase model using this augmented data;
- 4. At test time, apply the paraphraser to each reference in the test set; beam search is run for each of the n most probable sentence codes to produce n paraphrases per reference.

Output-guided constrained paraphrases

- For each sentence in a test set, we find all n-grams that are
  - (a) **not** in the **reference**
  - (b) but are **present** in at least 75% of the **system outputs**,
  - (c) limited to the top half of systems in the human system-level evaluation

Experiments

- **Goal**:
  - 1) evaluate the **adequacy** and **diversity** of our paraphrases;
  - 2) compare the usefulness of all methods in improving BLEU’s correlation with human judgments
- **Metric Evaluation**:
  - use Multi-ref BLEU and SENTBLEU to score all into-English system outputs
  - evaluate the scores by calculating the correlation with manual direct assessment (DA)

Paraphrase Adequacy

- Determine adequacy by manually evaluating paraphrases of the first 100 sentences of the de-en test set.
- 5 annotators rated the paraphrases' adequacy using DA, indicating how well (0-100) the official reference's meaning is preserved by its paraphrases.
- Baselines: BEAM, SAMPLED, LASER, TREELSTM, HUMAN.

Paraphrase Adequacy

Reference	DA	*What provoked Lindsay Lohan to such very strange actions is currently completely unclear.*	*Now they have come to an agreement.*
BEAM	91.7	What caused Lindsay Lohan to do such strange things is not clear at the moment. What provoked Lindsay Lohan's strange actions is not clear at the moment. What has provoked Lindsay Lohan's strange actions is not clear at the moment.	Now they've made a deal. Now they've reached a deal. Now they made a deal.
SAMPLED	85.0	What prompted Lindsay Lohan's most extraordinary actions? What made Lindsay Lohan act so weird? What inspired Lindsay Lohan to do such odd things?	And now they've agreed. And now they've agreed. They've reached an agreement.
LASER	90.1	What provoked Lindsay Lohan to act so strangely is not clear at the moment. It's not clear what provoked Lindsay Lohan to act so strangely. It's not clear what prompted Lindsay Lohan to act so strangely.	Now they've reached a deal. Now they've agreed. Now they've agreed
TREELSTM	88.0	What provoked Lindsay Lohan to do such a strange thing is not clear at the moment. It is not clear at this time what provoked Lindsay Lohan to do such strange things. The reason that Lindsay Lohan has been provoked by these very strange actions is not clear at the moment.	Now they made a deal. Now they've made a deal. They've already made a deal.
HUMAN	95.2	It is currently totally unclear what made Lindsay Lohan do such strange things. The cause of Lindsay Lohan's strange actions is really not clear at the moment. The reasons behind Lindsay Lohan's such bizarre acts are completely obscure for now.	They have now come to an agreement. An agreement has now been made. They have reached an agreement.

Table 1: Direct assessment (DA) adequacy scores for the BEAM and SAMPLED baseline, the two diverse approaches and human paraphrases for the 100-sentence de–en subset. We also provide each method's top 3 paraphrases for two references.

Paraphrase Diversity

- Evaluate the diversity of paraphrased references using two diversity scores (DS):

$$\mathrm{DS}_{\mathrm{x}} = \frac{1}{|Y|(|Y|-1)} \sum_{y \in Y} \sum_{y' \in Y, y' \neq y} 1 - \Delta_{\mathrm{x}}\left(y, y'\right),$$

- $\Delta_{\mathrm{x}}\left(y, y'\right)$ calculates the similarity of paraphrases y and y'. Two different functions are adopted: BOW (for lexical similarity) and TREE (for syntactic similarity).

Paraphrase Diversity

n	Method	DS_{BOW}	DS_{tree}	BLEU
0	none	-	-	29.8
5	RANDOM	0.10	0.01	34.8
	BEAM	0.22	0.30	37.0
	LASER	0.24	0.33	37.5
	TREELSTM	0.28	0.47	37.7
	SAMPLED	**0.41**	**0.56**	40.1
5*	SAMPLED	0.40	0.55	47.0
	Constraints	0.19	0.30	**56.5**
	HUMAN	**0.80**	**0.68**	48.9
20	RANDOM	0.10	0.01	34.8
	BEAM	0.27	0.37	39.7
	LASER	0.31	0.45	41.3
	TREELSTM	0.32	0.53	41.0
	SAMPLED	**0.51**	**0.65**	**47.3**
∞	Constraints	0.21	0.28	46.4
	MT submissions	0.37	0.51	-

Table 2: Diversity scores (DS) of paraphrased references averaged over all into-English test sets, where n is the number of paraphrases. The final row indicates diversity among MT outputs. * indicates results just for the 500-sentence de–en subset. The final column is the average BLEU score.

Metric Correlation Results

Approach	Method	System Gains Ave.	System Gains Min	System Gains Max	Segment Gains Ave.	Segment Gains Min	Segment Gains Max	System de–en	Segment de–en
Baselines (+5)	BEAM	0.020	-0.006	**0.059**	0.013	-0.001	0.029	**0.040**	0.021
	RANDOM	0.017	0.000	**0.046**	0.007	-0.002	0.017	**0.031**	0.017
	SAMPLED	0.024	-0.002	**0.067**	0.017	-0.004	**0.044**	**0.044**	0.043
Diversity (+1)	LASER	0.017	-0.000	**0.048**	0.009	-0.003	0.025	**0.034**	0.022
	TREELSTM	0.017	-0.000	**0.048**	0.011	-0.002	0.027	**0.031**	0.011
Diversity (+5)	LASER	0.020	-0.004	**0.056**	0.011	-0.002	0.033	**0.040**	0.022
	TREELSTM	0.020	-0.004	**0.057**	0.013	-0.004	0.030	**0.044**	0.008
Output-specific (+1)	LASER	0.012	-0.006	**0.041**	0.006	-0.001	0.016	**0.032**	0.015
	TREELSTM	0.014	-0.007	**0.041**	0.007	-0.005	0.016	**0.039**	0.011
Constraints	4-grams	0.025	-0.002	**0.061**	0.002	-0.097	**0.072**	-0.027	0.035
Human		-	-	-	-	-	-	**0.039**	0.037
WMT-19 best	Multiple	0.079	0.010	**0.194**	0.117	**0.072**	**0.145**	-	-

(a) Average and minimum and maximum gains over all into-English test sets (b) 500-sample subset

Table 3: Absolute gains in correlation (with respect to the true BLEU and sentenceBLEU baseline correlations). Significant gains (except for averages) are marked in bold ($p \leq 0.05$). Full results per language pair are provided in App. D. WMT-19 best refers to the best metric scores from the official shared task (the best metric can be different for each language pair).

Discussion

- **Does diversity help?**
  - The diversity of those paraphrases tends to positively correlate with gains in metric performance for both BLEU and SENT-BLEU.
  - The adequacy of the paraphrases appears to be a less important factor, shown by the fact that the best automatic diverse method at both levels was the SAMPLED baseline

n	Method	DS_{BOW}	DS_{tree}	BLEU
0	none	-	-	29.8
5	RANDOM	0.10	0.01	34.8
	BEAM	0.22	0.30	37.0
	LASER	0.24	0.33	37.5
	TREELSTM	0.28	0.47	37.7
	SAMPLED	**0.41**	**0.56**	40.1
5*	SAMPLED	0.40	0.55	47.0
	Constraints	0.19	0.30	**56.5**
	HUMAN	**0.80**	**0.68**	48.9
20	RANDOM	0.10	0.01	34.8
	BEAM	0.27	0.37	39.7
	LASER	0.31	0.45	41.3
	TREELSTM	0.32	0.53	41.0
	SAMPLED	**0.51**	**0.65**	**47.3**
∞	Constraints	0.21	0.28	46.4
	MT submissions	0.37	0.51	-

Table 2: Diversity scores (DS) of paraphrased references averaged over all into-English test sets, where n is the number of paraphrases. The final row indicates diversity among MT outputs. * indicates results just for the 500-sentence de–en subset. The final column is the average BLEU score.

Discussion

- **What effect do more references have?**
  - Diversity is positively correlated with gains for most language directions, however improvements are slight.
  - The initial paraphrase has the most impact…

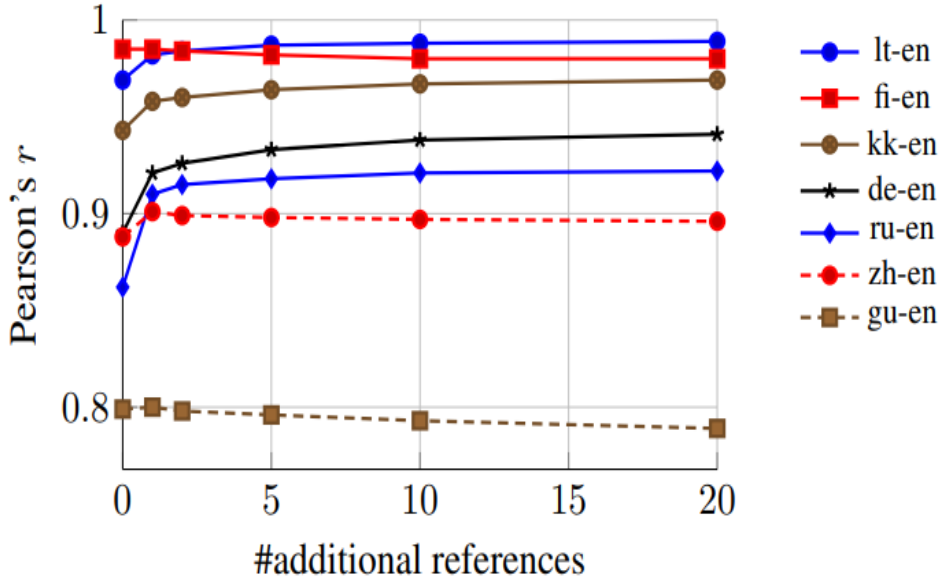

Figure 2: TREELSTM system-level correlations (+0-20).

Discussion

- **Why are gains only slight?**
  - Although all the systems improve a fair number of comparisons, they degrade almost as many.
  - The same pattern can be seen for human paraphrases: 6.46% being degraded vs. 8.30% improved
  - BLEU is a balancing act …

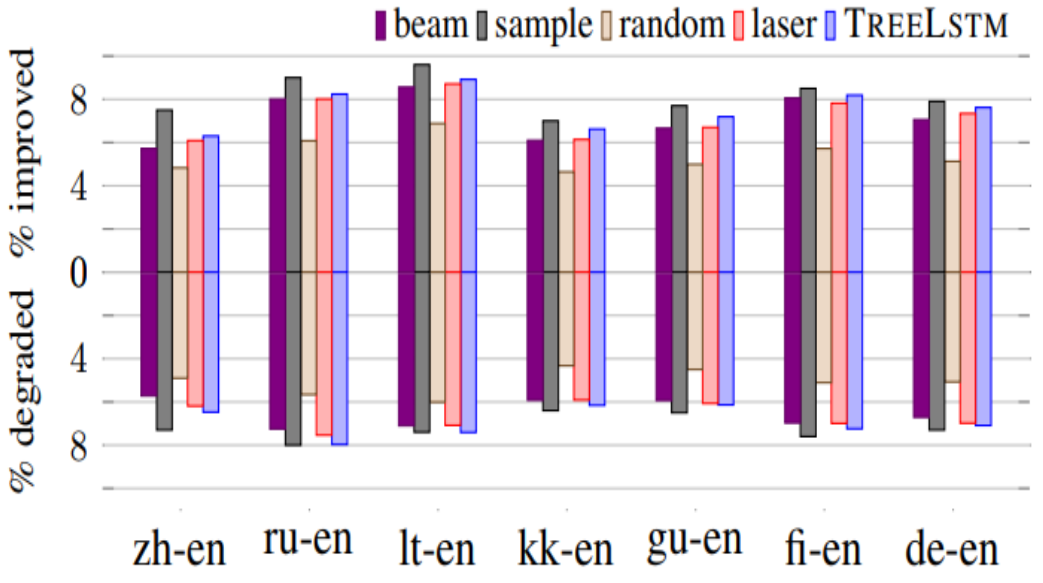

Figure 3: % improved and degraded (with respect to single-reference sentence-BLEU) for methods with +5 references.

Discussion

- **What is the effect on individual n-grams?**

N	newly matched ngrams	missing ngrams
1	a (494) of (480) , (442) to (370) in (364) The (315) the (273) is (204) for (196) has (196) on (193) was (179) have (171) that (166) be (155) at (145) been (140) with (138) and (134)	to (921) in (921) on (870) is (802) of (798) a (786) for (568) The (556) with (509) it (508) has (505) are (482) by (480) was (478) have (449) - (443) at (437) as (426) which (386)
4	U . S . (63) the U . S (39) , as well as (19) p . m . (15) for the first time (13) in accordance with the (12) the United States , (11) in the United States (10) a member of the (10) of the United States (9) The U . S (9) . m . on (9) , in order to (9) the United States and (8) , of course , (8) . S . Navy (8) . m . , (8) the Chinese Academy of (8) Chinese Academy of Engineering (8) the renaming of the (7)	U . S . (136) , according to the (99) , ” he said (77) the U . S (55) of the United States (48) of the Ministry of (39) the end of the (38) , ” said the (37) same time , the (36) , such as the (36) as well as the (35) (Xinhua) – (34) and so on . (33) , he said . (32) the head of the (32) , the head of (31) , as well as (30) on the basis of (30) , and so on (29)

Table 4: Most frequently newly matched and missing n-grams for the de–en and ru–en test sets for BEAM (+5).

Conclusion

- Experiments show that adding paraphrased references rarely hurts BLEU and can provide moderate gains in its correlation with human judgments.
- Manual paraphrasing does give the best system-level BLEU results, but these gains are relatively limited, suggesting that diversity alone has its limits in addressing weakness of surface-based evaluation metrics like BLEU.

Introduction

- Open-domain dialogue datasets with only a **single relevant response and no irrelevant responses** are not suitable for training and testing dialog evaluation models.
- Irrelevant responses can easily be generated by **sampling random utterances from other contexts**, but such examples typically do not have any overlap with the context and hence are **easier for the model to distinguish** from relevant responses.

Contributions

- propose a multi-reference open-domain dialogue dataset with multiple relevant responses and adversarial irrelevant responses.
- perform an extensive study of the existing dialogue evaluation metrics using this dataset
- propose a new transformer-based evaluator pretrained on large-scale dialogue datasets.

Proposed Dataset

- Additional 5 reference responses were collected with the help of human annotators for each of the 19k contexts derived from DailyDialog.
- Human annotators were also asked to carefully irrelevant responses that have a significant word overlap with the context.

Total # of contexts	19,071
Avg. # of turns per context	3.31
Avg. # of words per context	45.32
Avg. # of words per utterance	13.55
# of contexts with 5 relevant responses	19,071
# of contexts with 5 adv. irrelevant responses	11,429
Avg. # of words per relevant response	10.13
Avg. # of words per irrelevant response	13.8

Table 2: DailyDialog++ dataset statistics.

Example

Context	Valid responses	Invalid, adversarial responses
FS: Can you do push-ups ?	SS: You don't believe me, do you?	SS: Push up the window and look out for a minute
SS: Of course I can . It's a piece of cake !	SS: Start your timer, here we go.	SS: Would you like to eat a piece of cake before gym?
Believe it or not , I can do 30 push-ups a minute.	SS: Watch me do it.	SS: I like watching the Ripley's Believe it or Not show
FS: Really ? I think that's impossible !	SS: That's because you can't do it.	where they discuss nearly impossible feats
SS: You mean 30 push-ups ?	SS: You don't know that I am a	and gymnastics
FS: Yeah !	fitness trainer, do you ?	SS: I have enough time for my treadmill exercises
		SS: Are you asking me to do 40 squats?

Table 1: Examples from DailyDialog++ dataset with the context consisting of 2 speakers [annotated as FS (First Speaker) and SS (Second Speaker)], and multiple reference responses and adversarial negative responses. The underlined, purple colored words in the adversarial responses are those that overlap or are closely related to the theme or words in the context.

Dialogue Evaluation using BERT

- Existing BERT-based evaluation metrics do not leverage a successful recipe of (i) pretraining with a masked language modeling objective and (ii) finetuning with a task-specific objective
- DEB is trained using a masked language model objective (similar to BERT) and a modified next response prediction objective (identifying whether the given response is a valid next response for the given context)
- The key contribution here is to assess if pretraining on large-scale dialogue corpora improves the performance of dialogue evaluation metrics.

Experimental Setup

- The goal is to check if the **adversarial responses** in the proposed dataset, which are specifically crafted to **target context-dependent** model-based metrics, indeed affect the performance of such models.
- To do so….
  - 1) first benchmark the models' performance on random negatives
  - 2) then check if the performance drops when evaluated on adversarial examples

Experimental Setup

- For each context in the test set, we obtain the scores assigned by a given metric to the 5 positive and 5 negative responses.
- For all **untrained metrics** (e.g BLEU), we consider the **remaining 4 relevant responses** as references.

Performance on Random Negatives

- The performance of all metrics is quantified using two measures:
  - 1) Point Biserial Correlation (PBC) between the scores assigned by a metric and the binary target (1 for positive example and 0 for negative example)
  - 2) classification accuracy of the metric by using a threshold and marking all responses having a score above this threshold as positive and others as negative. (0.5 for trained metrics and for untrained metrics, they perform a search from 0 to 1 with step size of 0.001 and select the threshold that minimizes the error rate on valid set)

Performance on Random Negatives

Metric	Point Biserial Correlation (p-value)				Accuracy in percentage			
	Single	Multiple			Single	Multiple		
		Avg	Max	Standard		Avg	Max	Standard
BLEU-1	0.26 (<1e-9)	0.42 (<1e-9)	0.41 (<1e-9)	0.41 (<1e-9)	61.26	68.60	68.75	70.36
BLEU-2	0.22 (<1e-9)	0.39 (<1e-9)	0.36 (<1e-9)	0.40 (<1e-9)	58.09	68.26	68.37	68.66
BLEU-3	0.14 (<1e-9)	0.26 (<1e-9)	0.24 (<1e-9)	0.28 (<1e-9)	53.11	58.85	58.90	58.89
BLEU-4	0.08 (<1e-9)	0.17 (<1e-9)	0.15 (<1e-9)	0.18 (<1e-9)	51.16	53.56	53.56	53.50
METEOR	0.23 (<1e-9)	0.40 (<1e-9)	0.41 (<1e-9)	–	59.77	68.51	68.01	–
ROUGE-L	0.23 (<1e-9)	0.41 (<1e-9)	0.40 (<1e-9)	0.37 (<1e-9)	59.47	67.89	68.25	68.43
deltaBLEU (Galley et al., 2015)	–	–	–	0.29 (<1e-9)	–	–	–	64.89
Embed Avg	0.23 (<1e-9)	0.25 (<1e-9)	0.23 (<1e-9)	–	61.27	61.56	62.67	–
Vec Extr (Forgues et al., 2014)	0.24 (<1e-9)	0.35 (<1e-9)	0.33 (<1e-9)	–	59.22	63.70	63.90	–
GreedyMatch (Rus and Lintean, 2012)	0.24 (<1e-9)	0.36 (<1e-9)	0.32 (<1e-9)	–	60.02	63.99	65.56	–
BERTScore (Zhang et al., 2020a)	0.29 (<1e-9)	0.39 (<1e-9)	0.39 (<1e-9)	–	63.71	69.05	68.59	–
ADEM (Lowe et al., 2017)	0.40 (<1e-9)				64.74			
BERT regressor (Shimanaka et al., 2019)	0.52 (<1e-9)				73.40			
BERT+DNN (Ghazarian et al., 2019)	0.57 (<1e-9)				74.67			
RUBER (Tao et al., 2018)	0.64 (<1e-9)				78.18			
RUBER-Large (Tao et al., 2018)	0.69 (<1e-9)				82.36			
DEB (ours)	**0.79*** (<1e-9)				**88.27***			

Table 3: Automatic evaluation metrics performance on random negatives (PBC refers to point-biserial correlation. Column subheading 'Single' refers to experiments using single reference response and 'Avg' and 'Max' are the average and maximum aggregation strategies when using multiple reference responses. 'Standard' is applicable when the metric aggregates multiple references differently. * indicates statistical significance in performance over all other metrics (with p-values <1e-9) on William's test for comparing correlations and Chi-squared test for accuracies. p-values for individual correlations are in parenthesis.

Analysis using Box Plots

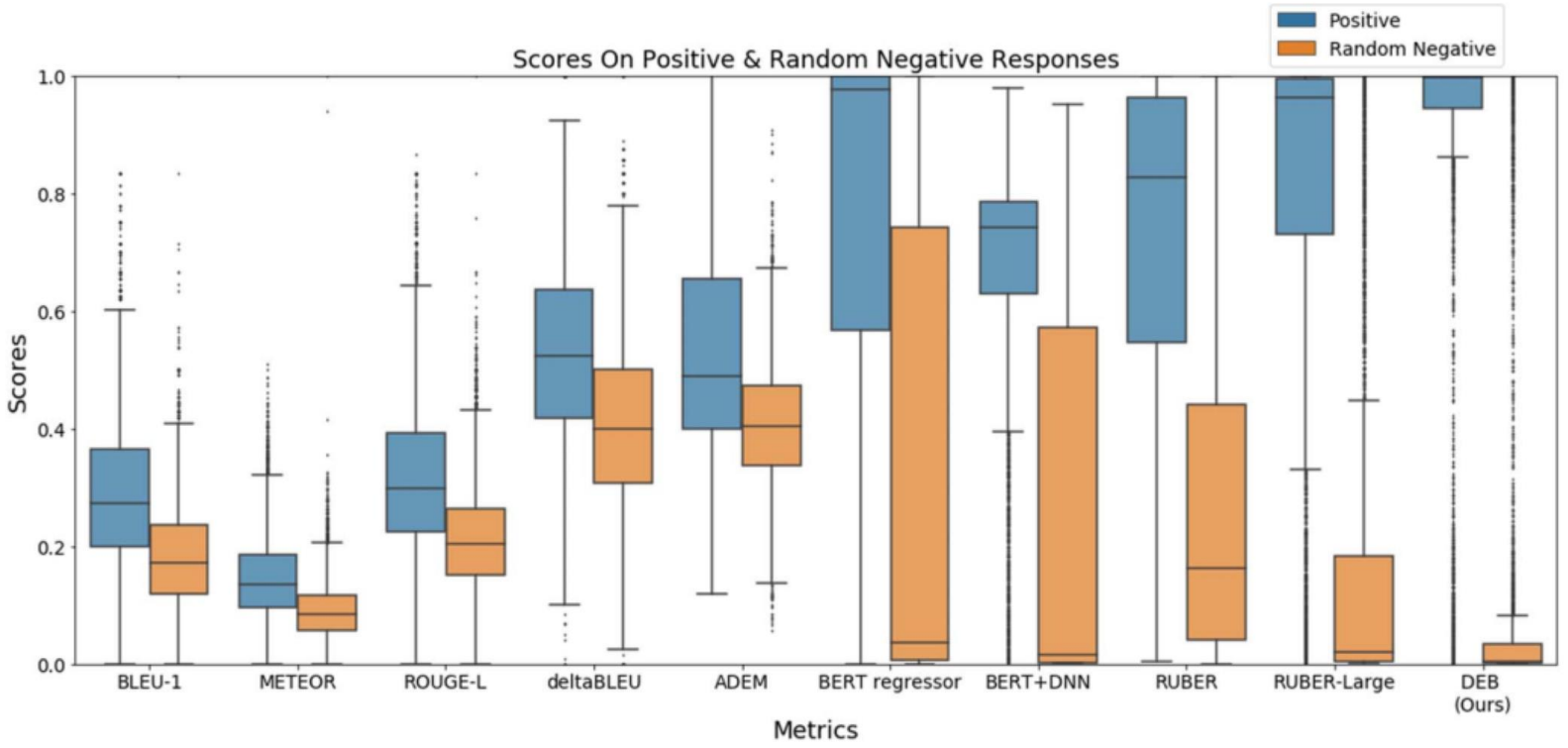

Figure 1: Box plots of the scores given by various metrics to the positive and random negative responses.

Performance on Synthetically Crafted Adversarial Responses

- Before evaluating them using the adversarial examples, we first investigate the performance of the models with **synthetically crafted** adversarial attacks.
- Perform a simple transformations on relevant responses by:
  - 1) jumbling words in the sequence
  - 2) reversing the sequence
  - 3) dropping all words except nouns
  - 4) dropping all stopwords
  - 5)dropping punctuation
  - 6) replacing words with synonyms

Performance on Synthetically Crafted Adversarial Responses

Modification	DEB	RUBER-Large	RUBER	BERT+DNN
	% classified as positive			
Unmodified positives	87.9%	81.7%	77.5%	93.5%
Reverse word order	60.0%	70.3%	71.3%	80.4%
Jumble word order	69.3%	71.2%	72.3%	77.4%
Retain only nouns	60.1%	27.9%	27.8%	0.0%
Remove punctuation	86.4%	72.9%	72.4%	88.5%
Remove stopwords	85.8%	73.6%	69.6%	29.3%
Replace with synonyms	81.2%	70.8%	65.6%	91.1%
	Pearson Correlation with human scores			
Remove stopwords	0.58 (<1e-9)	0.56 (<1e-9)	0.52 (<1e-9)	0.056 (0.26)
Replace with synonyms	0.68 (<1e-9)	0.57 (<1e-9)	0.54 (<1e-9)	−0.017 (0.67)

Table 4: Fraction of responses classified as positives with synthetic modifications. Unmodified positives are presented in the 1st row for reference (p-values for individual correlations in brackets).

Performance of Model-Based Metrics on Manually Crafted Adversarial Responses

- The accuracy of all the model drops.
- The models wrongly classify most of the irrelevant responses as positive responses.

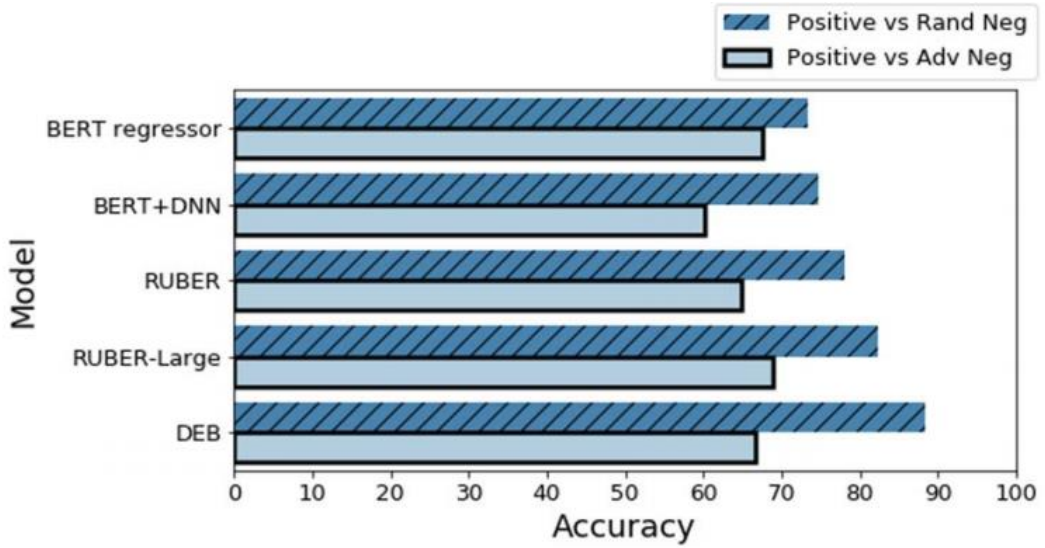

Figure 2: Accuracy of different models in identifying adversarial and random negatives versus positive responses.

Conclusion

- Even in the presence of multiple correct references, n-gram based metrics and embedding based metrics do not perform well at separating relevant responses from even random negatives.
- While model-based metrics perform better than n-gram and embedding based metrics on random negatives, their performance drops substantially when evaluated on adversarial examples.
- Even large-scale pretrained evaluation models are not robust to the adversarial examples in the dataset.