https://arxiv.org/pdf/2004.04696.pdf

Paper review

BLEURT: Learning Robust Metrics for Text Generation (ACL 2020)

Presentation: **Jeiyoon Park**
6th Generation, TAVE

TAVE Research

Outline

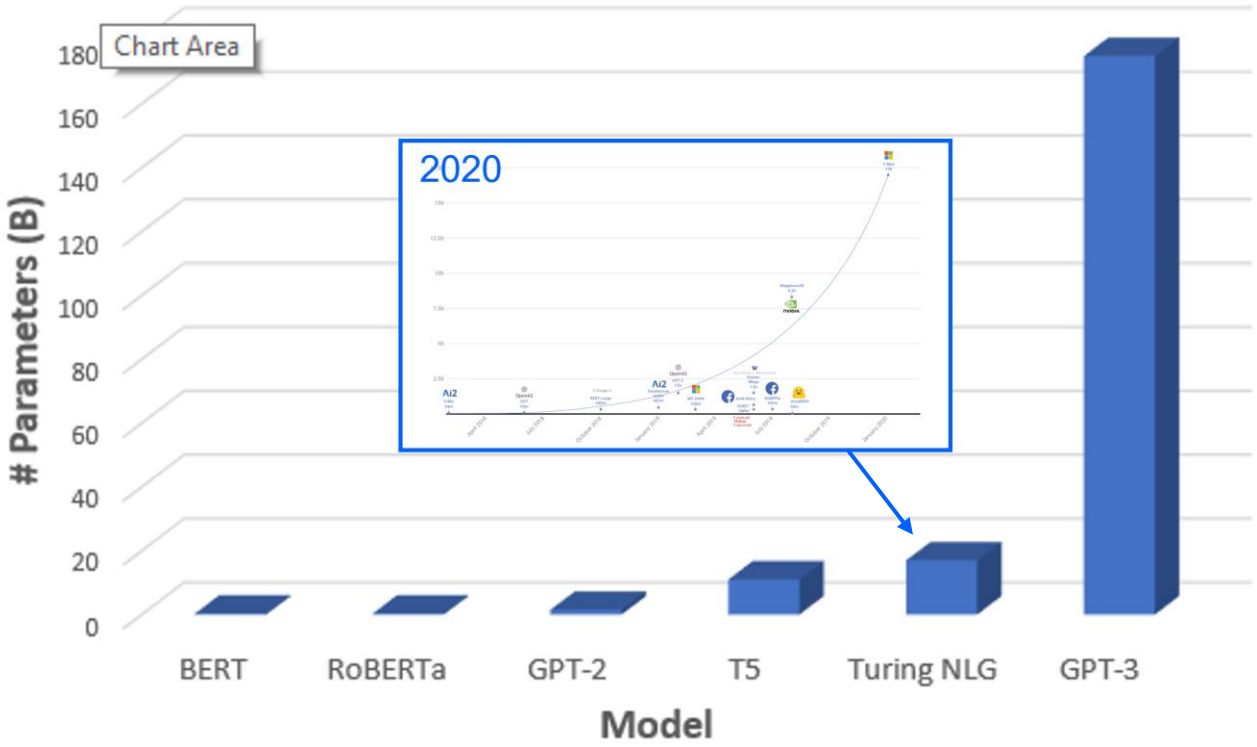
1. Contribution
2. Method
3. Experiments
4. Conclusion

Outline

1. Contribution
2. Method
3. Experiments
4. Conclusion

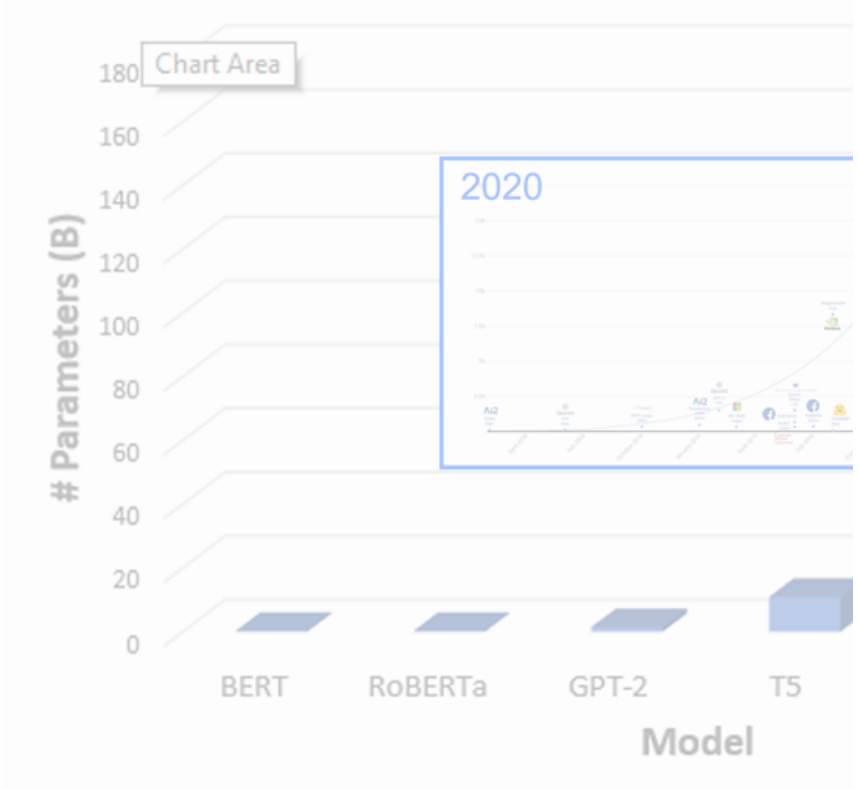
Contribution

1. Automatic Evaluation



Contribution

1. Automatic Evaluation



Measurement Tools

Contribution

1. Automatic Evaluation

1) Designing crowd sourcing experiments is an expensive and high-latency process

2) So, NLG researchers commonly employ “automatic evaluation metrics” (e.g., BLEU and ROUGE)

3) However, theses metrics rely on N-gram overlap which are only sensitive to lexical variation

4) In “IID assumption”, automatic evaluation is problematic because of domain drifts, and quality drifts: a model trained on ratings data from 2015 may fail to distinguish top performing systems in 2019

Contribution

2. BLEURT

1) It is based on *BERT (Devlin et al., 2019)*

2) BLEURT is a novel pre-training scheme, which uses random perturbations of Wikipedia sentences, augmented with a diverse set of lexical and semantic-level supervision signals

3) This paper investigates sentence-level, reference based metrics, which describe the extent to which a candidate sentence is similar to a reference one.

4) The exact definition of similarity may range from string overlap to logical entailment.

Contribution

3. Experiments

1) This paper first verifies that it provides state-of-the-art results on all recent years of the WMT Metrics Shared task (2017 to 2019, to-English language pairs)

2) And then stress-test its ability to cope with quality drifts with a synthetic benchmark based on WMT 2017.

3) Finally, They show that it can easily adapt to a different domain with three tasks from a data-to-text dataset, *WebNLG 2017 (Garden et al., 2017)*

Outline

1. Contribution
2. Method
3. Experiments
4. Conclusion

Preliminaries

1. Notation

$x = (x_1, \ldots, x_r)$: Reference sentence, where r is length
$\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_p)$: Prediction sentence, where p is length

$\{(x_i, \tilde{x}_i, y_i)\}_{n=1}^{N}$: training dataset

, where N is size of training dataset, and y_i is Human rating that indicates how good $\tilde{x}_i$ is w.r.t x_i

2. Goal

To learn a function $f : (x, \tilde{x}) \rightarrow y$

Method

1. Fine-Tuning BERT for Quality Evaluation

- Given x and $\tilde{x}$, BERT returns a sequence of contextualized vectors:

$$v_{[CLS]}, v_{x_1}, \dots, v_{x_r}, v_1, \dots, v_{\tilde{x}_p} = BERT(x, \tilde{x})$$

, where $v_{[CLS]}$ is the representation for the special $[CLS]$

- BERT adds a linear layer on top of the $[CLS]$ vector to predict the rating:

$$\hat{y} = f(x, \tilde{x}) = W\tilde{v}_{[CLS]} + b$$

, where W and b are the weight matrix and bias vector respectively

Method

2. Pre-Training on Synthetic Data

1) This method is a pre-training technique that they use to "warm up" BERT before fine-tuning on rating data *(Note: It's just addition, not a replacement)*

2) They generate a large # of synthetic reference-candidate pairs $(z, \tilde{z})$, and train BERT on several lexical-level and semantic-level supervision signals with a multi-task loss

3) Three requirements

- The set of reference sentences should be large and diverse
- The sentence pairs should contain a wide variety of lexical, syntactic, and semantic dissimilarities
- The pre-training objectives should effectively capture those phenomena

Method: Pre-training on Synthetic Data

1. Generating Sentence Pairs

- Existing datasets may fail to capture the errors and alteration (e.g., omission, repetitions, non-sensical substitutions)

- This paper generates synthetic sentence pairs $(z, \tilde{z})$ by randomly perturbing 1.8 M segments from Wikipedia

- BLEURT employs three techniques: (1) mask-filling with BERT, (2) backtranslation, and (3) randomly dropping out words

- They obtain 6.5 M perturbations $\tilde{z}$

Method: Pre-training on Synthetic Data

1. Generating Sentence Pairs

1) Mask-filling with BERT

- BLEURT inserts masks at random positions in the Wikipedia sentences and fill them with the language model

- BLEURT introduce lexical alterations while maintaining the fluency of the sentence

- 15 masks per sentence (Both sampling random words and creating contiguous sequence)

Method: Pre-training on Synthetic Data

1. Generating Sentence Pairs

2) Backtranslation

- BLEURT generates paraphrases and perturbations with backtranslation

- This method aims to create variants of the reference sentence that serves semantics

Method: Pre-training on Synthetic Data

1. Generating Sentence Pairs

3) Dropping words

- They found it useful in their experiments to randomly drop words from the synthetic examples above to create other examples

Method: Pre-training on Synthetic Data

2. Pre-Training Signals

- Each sentence pair $(z, \tilde{z})$ is augmented with a set of pre-training signal $\{\tau_k\}$, where τ_k is the target vector of pre-training task k

- Good pre-training signals should capture a wide variety of lexical and semantic differences

- $\{\tau_k\}$ is cheap to obtain, so that the approach can scale to large amount of synthetic data

Method: Pre-training on Synthetic Data

2. Pre-Training Signals

1) Automatic Metrics

- BLEURT creates three signals $\tau_{BLEU}, \tau_{ROUGE}, \tau_{BERT_{score}}$ with sentence BLEU, ROUGE, and BERTscore respectively

Method: Pre-training on Synthetic Data

2. Pre-Training Signals

2) Backtranslation Likelihood

- To leverage existing translation models to measure semantic equivalence

- Given a reference-candidate pair $(z, \tilde{z})$, the training signal measures the probability that $\tilde{z}$ is a backtranslation of z, $p(\tilde{z}|z)$, normalized by the length of $\tilde{z}$

e.g.) $P_{en \to fr}(z_{fr}|z)$: a translation model that assigns probabilities to French sentence z_{fr} conditioned on English sentences z:

$$\tau_{en-fr,\tilde{z}|z} = \frac{logP(\tilde{z}|z)}{|\tilde{z}|}, where\ |\tilde{z}|\ is\ \#\ of\ tokens\ in\ \tilde{z}$$

- $\tau_{en-fr,\tilde{z}|z}$, $\tau_{en-fr,z|\tilde{z}}$, $\tau_{en-de,\tilde{z}|z}$, $\tau_{en-de,z|\tilde{z}}$ (four signals for pre-training)

Method: Pre-training on Synthetic Data

2. Pre-Training Signals

$\mathbf{P}^a$	A senior is waiting at the window of a restaurant that serves sandwiches.	Relationship
$\mathbf{H}^b$	A person waits to be served his food.	Entailment
	A man is looking to order a grilled cheese sandwich.	Neutral
	A man is waiting in line for the bus.	Contradiction

[a] P, Premise.
[b] H, Hypothesis.

3) Text Entailment

- τ_{entail} express whether z entails or contradicts $\tilde{z}$ using classifier

4) Backtranslation flag

- $\tau_{backtrans\_flag}$ is a Boolean that indicates whether the perturbation was generated with backtranslation or with mask-filling

Task Type	Pre-training Signals	Loss Type
BLEU	τ_{BLEU}	Regression
ROUGE	$\tau_{\text{ROUGE}} = (\tau_{\text{ROUGE-P}}, \tau_{\text{ROUGE-R}}, \tau_{\text{ROUGE-F}})$	Regression
BERTscore	$\tau_{\text{BERTscore}} = (\tau_{\text{BERTscore-P}}, \tau_{\text{BERTscore-R}}, \tau_{\text{BERTscore-F}})$	Regression
Backtrans. likelihood	$\tau_{\text{en-fr},z\|\tilde{z}}, \tau_{\text{en-fr},\tilde{z}\|z}, \tau_{\text{en-de},z\|\tilde{z}}, \tau_{\text{en-de},\tilde{z}\|z}$	Regression
Entailment	$\tau_{\text{entail}} = (\tau_{\text{Entail}}, \tau_{\text{Contradict}}, \tau_{\text{Neutral}})$	Multiclass
Backtrans. flag	$\tau_{\text{backtran\_flag}}$	Multiclass

Method: Pre-training on Synthetic Data

3. Modeling

- For each pre-training task k, BLEURT uses either a regression or classification loss

- And then aggregate the task-level losses with a weighted sum

- Let τ_k describe the target vector for each task (e.g., the probability for the class *Entail, Contradict, Neutral*, or the precision, recall, and F-score for ROUGE)

- They define their aggregate pre-training loss function as follows:

$\ell_{pre-training} = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} r_k \ell_k(\tau_k^m, \widehat{\tau_k^m})$, where τ_k^m is the target vector for example m, M is # of synthetic examples, r_k are hyperparameter weights (obtained with grid search)

Outline

1. Contribution
2. Method
3. Experiments
4. Conclusion

Experiments

1. Two tasks: translation & data-to-text

- First, they benchmarks BLEURT against existing text generation metrics on the last three years of the WMT Metrics Shared Task

- They then evaluates its robustness to quality drifts with a series of synthetic datasets based on WMT17

- They tests BLEURT's ability to adapt to different tasks with the WebNLG 2017 Challenge Dataset

Experiments

2. Three steps of training BLEURT

1) Regular BERT pre-training
2) pre-training on synthetic data
3) fine-tuning on task-specific ratings

3. Two versions of BLEURT

1) *BLEURT*
2) *BLEURTbase*

- Respectively based on BERT-Large (24 layers, 1024 hidden units, 16 heads), and BERT-Base (12 layers, 768 hidden units, 12 heads)

Experiments: (1) WMT Metrics Shared Task

1. Datasets

- 2017 to 2019 of the WMT Metrics Shared Task, to-English language pairs

- Official WMT test set, which include several thousand pairs of sentences with human ratings from the news domain

- The training sets contain 5,360, 9,492, and 147,691 records for each year.

- The test sets for years 2018 and 2019 are noisier

Experiments: (1) WMT Metrics Shared Task

2. Metrics

- Agreement between the automatic metrics and the human ratings

- Kendall's Tau τ, for consistency across experiments

- the official WMT metric for that year, for completeness

- The official WMT metric is either Pearson's correlation or a robust variant of Kendall's Tau called DARR

Experiments: (1) WMT Metrics Shared Task

3. Results: WMT17 Metrics Shared Task

model	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	**avg**
	τ / r	τ / r	τ / r	τ / r	τ / r	τ / r	τ / r	τ / r
sentBLEU	29.6 / 43.2	28.9 / 42.2	38.6 / 56.0	23.9 / 38.2	34.3 / 47.7	34.3 / 54.0	37.4 / 51.3	32.4 / 47.5
MoverScore	47.6 / 67.0	51.2 / 70.8	NA	NA	53.4 / 73.8	56.1 / 76.2	53.1 / 74.4	52.3 / 72.4
BERTscore w/ BERT	48.0 / 66.6	50.3 / 70.1	61.4 / 81.4	51.6 / 72.3	53.7 / 73.0	55.6 / 76.0	52.2 / 73.1	53.3 / 73.2
BERTscore w/ roBERTa	54.2 / 72.6	56.9 / 76.0	64.8 / 83.2	56.2 / 75.7	57.2 / 75.2	57.9 / 76.1	58.8 / 78.9	58.0 / 76.8
chrF++	35.0 / 52.3	36.5 / 53.4	47.5 / 67.8	33.3 / 52.0	41.5 / 58.8	43.2 / 61.4	40.5 / 59.3	39.6 / 57.9
BEER	34.0 / 51.1	36.1 / 53.0	48.3 / 68.1	32.8 / 51.5	40.2 / 57.7	42.8 / 60.0	39.5 / 58.2	39.1 / 57.1
BLEURTbase -pre	51.5 / 68.2	52.0 / 70.7	66.6 / 85.1	60.8 / 80.5	57.5 / 77.7	56.9 / 76.0	52.1 / 72.1	56.8 / 75.8
BLEURTbase	55.7 / 73.4	56.3 / 75.7	68.0 / 86.8	**64.7** / 83.3	60.1 / 80.1	62.4 / 81.7	59.5 / 80.5	61.0 / 80.2
BLEURT -pre	56.0 / 74.7	57.1 / 75.7	67.2 / 86.1	62.3 / 81.7	58.4 / 78.3	61.6 / 81.4	55.9 / 76.5	59.8 / 79.2
BLEURT	**59.3 / 77.3**	**59.9 / 79.2**	**69.5 / 87.8**	64.4 / **83.5**	**61.3 / 81.1**	**62.9 / 82.4**	**60.2 / 81.4**	**62.5 / 81.8**

Table 2: Agreement with human ratings on the WMT17 Metrics Shared Task. The metrics are Kendall Tau (τ) and the Pearson correlation (r, the official metric of the shared task), divided by 100.

Experiments: (1) WMT Metrics Shared Task

3. Results: WMT18 Metrics Shared Task

model	cs-en τ / DA	de-en τ / DA	et-en τ / DA	fi-en τ / DA	ru-en τ / DA	tr-en τ / DA	zh-en τ / DA	**avg** τ / DA
sentBLEU	20.0 / 22.5	31.6 / 41.5	26.0 / 28.2	17.1 / 15.6	20.5 / 22.4	22.9 / 13.6	21.6 / 17.6	22.8 / 23.2
BERTscore w/ BERT	29.5 / 40.0	39.9 / 53.8	34.7 / 39.0	26.0 / 29.7	27.8 / 34.7	31.7 / 27.5	27.5 / 25.2	31.0 / 35.7
BERTscore w/ roBERTa	31.2 / 41.1	42.2 / 55.5	37.0 / 40.3	27.8 / 30.8	30.2 / 35.4	32.8 / 30.2	29.2 / 26.3	32.9 / 37.1
Meteor++	22.4 / 26.8	34.7 / 45.7	29.7 / 32.9	21.6 / 20.6	22.8 / 25.3	27.3 / 20.4	23.6 / 17.5*	26.0 / 27.0
RUSE	27.0 / 34.5	36.1 / 49.8	32.9 / 36.8	25.5 / 27.5	25.0 / 31.1	29.1 / 25.9	24.6 / 21.5*	28.6 / 32.4
YiSi1	23.5 / 31.7	35.5 / 48.8	30.2 / 35.1	21.5 / 23.1	23.3 / 30.0	26.8 / 23.4	23.1 / 20.9	26.3 / 30.4
YiSi1 SRL 18	23.3 / 31.5	34.3 / 48.3	29.8 / 34.5	21.2 / 23.7	22.6 / 30.6	26.1 / 23.3	22.9 / 20.7	25.7 / 30.4
BLEURTbase -pre	33.0 / 39.0	41.5 / 54.6	38.2 / 39.6	30.7 / 31.1	30.7 / 34.9	32.9 / 29.8	28.3 / 25.6	33.6 / 36.4
BLEURTbase	34.5 / **42.9**	43.5 / 55.6	39.2 / 40.5	31.5 / 30.9	31.0 / 35.7	35.0 / 29.4	29.6 / **26.9**	34.9 / 37.4
BLEURT -pre	34.5 / 42.1	42.7 / 55.4	39.2 / 40.6	31.4 / 31.6	31.4 / 34.2	33.4 / 29.3	28.9 / 25.6	34.5 / 37.0
BLEURT	**35.6** / 42.3	**44.2 / 56.7**	**40.0 / 41.4**	**32.1 / 32.5**	**31.9 / 36.0**	**35.5 / 31.5**	**29.7** / 26.0	**35.6 / 38.1**

Table 3: Agreement with human ratings on the WMT18 Metrics Shared Task. The metrics are Kendall Tau (τ) and WMT's Direct Assessment metrics divided by 100. The star * indicates results that are more than 0.2 percentage points away from the official WMT results (up to 0.4 percentage points away).

Experiments: (1) WMT Metrics Shared Task

3. Results: WMT19 Metrics Shared Task

model	de-en τ / DA	fi-en τ / DA	gu-en τ / DA	kk-en τ / DA	lt-en τ / DA	ru-en τ / DA	zh-en τ / DA	**avg** τ / DA
sentBLEU	19.4 / 5.4	20.6 / 23.3	17.3 / 18.9	30.0 / 37.6	23.8 / 26.2	19.4 / 12.4	28.7 / 32.2	22.7 / 22.3
BERTscore w/ BERT	26.2 / 17.3	27.6 / 34.7	25.8 / 29.3	36.9 / 44.0	30.8 / 37.4	25.2 / 20.6	37.5 / 41.4	30.0 / 32.1
BERTscore w/ roBERTa	29.1 / 19.3	29.7 / 35.3	27.7 / **32.4**	37.1 / 43.1	32.6 / 38.2	26.3 / **22.7**	41.4 / **43.8**	32.0 / **33.6**
ESIM	28.4 / 16.6	28.9 / 33.7	27.1 / 30.4	38.4 / 43.3	33.2 / 35.9	26.6 / 19.9	38.7 / 39.6	31.6 / 31.3
YiSi1 SRL 19	26.3 / **19.8**	27.8 / 34.6	26.6 / 30.6	36.9 / 44.1	30.9 / 38.0	25.3 / 22.0	38.9 / 43.1	30.4 / 33.2
BLEURTbase -pre	30.1 / 15.8	30.4 / 35.4	26.8 / 29.7	37.8 / 41.8	34.2 / 39.0	27.0 / 20.7	40.1 / 39.8	32.3 / 31.7
BLEURTbase	31.0 / 16.6	31.3 / 36.2	27.9 / 30.6	**39.5** / **44.6**	**35.2** / 39.4	**28.5** / 21.5	41.7 / 41.6	33.6 / 32.9
BLEURT -pre	31.1 / 16.9	31.3 / **36.5**	27.6 / 31.3	38.4 / 42.8	35.0 / 40.0	27.5 / 21.4	41.6 / 41.4	33.2 / 32.9
BLEURT	**31.2** / 16.9	**31.7** / 36.3	**28.3** / 31.9	**39.5** / **44.6**	**35.2** / **40.6**	28.3 / 22.3	**42.7** / 42.4	**33.8** / **33.6**

Table 4: Agreement with human ratings on the WMT19 Metrics Shared Task. The metrics are Kendall Tau (τ) and WMT's Direct Assessment metrics divided by 100. All the values reported for Yisi1_SRL and ESIM fall within 0.2 percentage of the official WMT results.

Experiments: (2) Robustness to Quality Drift

3. Results: WMT17

- They creates increasingly challenging datasets by sub-sampling the records from the WMT Metrics shared task, keeping low-rated translations for training and high-rated translations for test

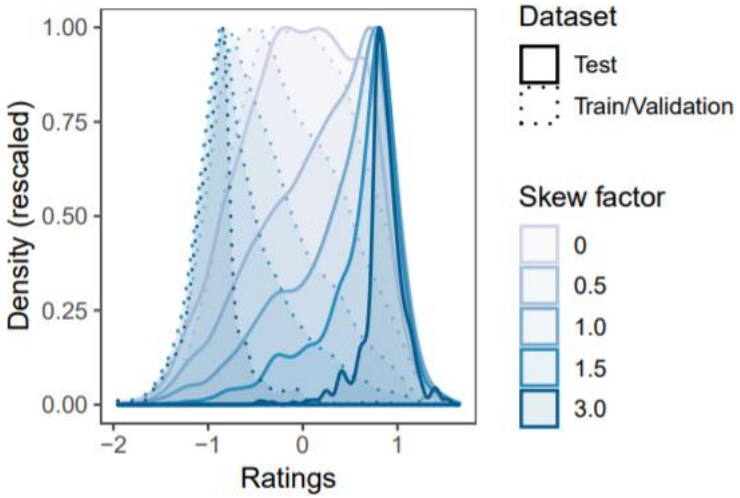


Figure 1: Distribution of the human ratings in the train/validation and test datasets for different skew factors.

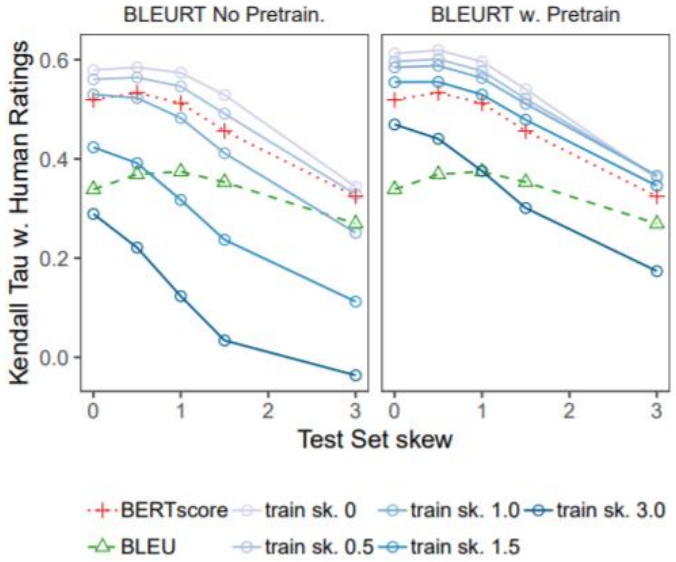


Figure 2: Agreement between BLEURT and human ratings for different skew factors in train and test.

Experiments: (3) WebNLG Experiments

1. Dataset: WebNLG Challenge 2017

- The WebNLG challenge benchmarks systems that produce natural language description of entities (e.g., buildings, cities, artists)

- The submissions are evaluated on 3 aspects: semantics, grammar, and fluency

Experiments: (3) WebNLG Experiments

3. Results: WebNLG Challenge 2017

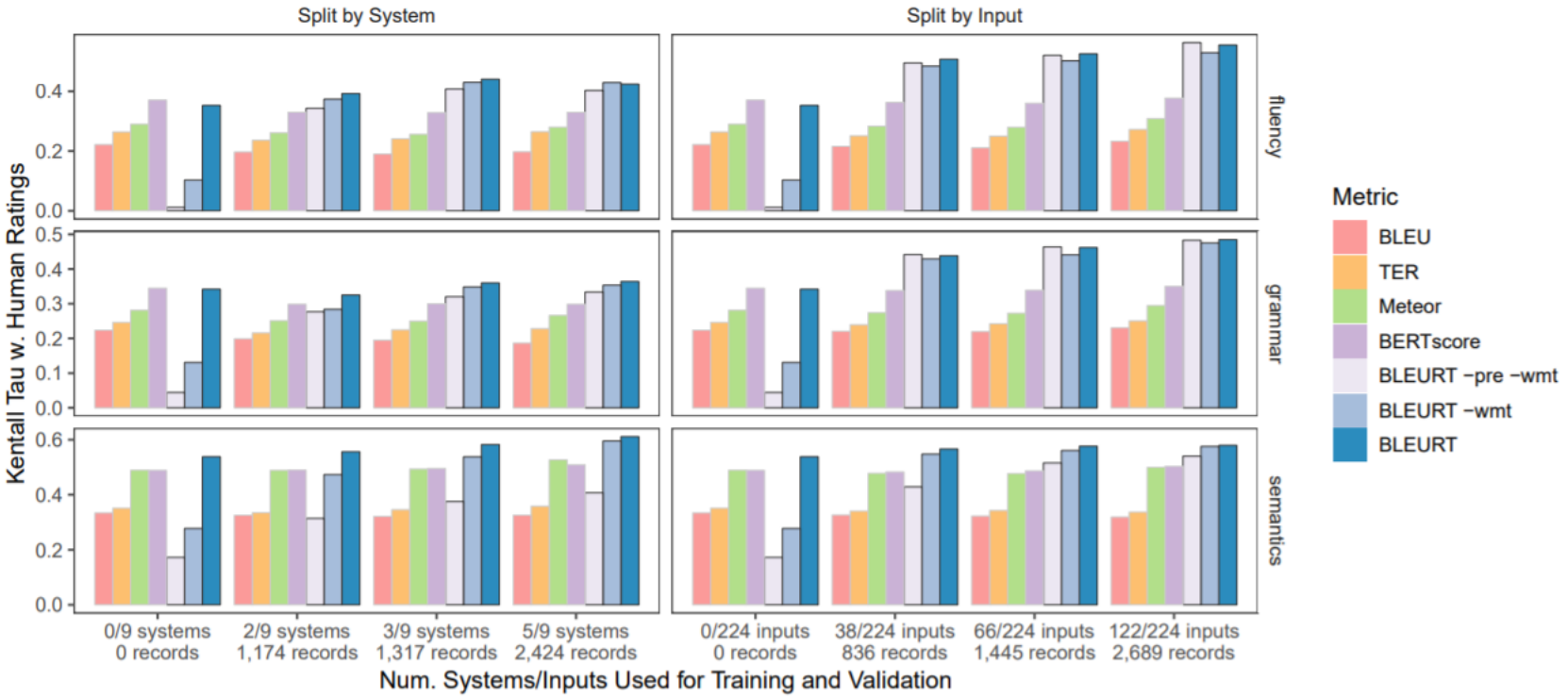


Figure 3: Absolute Kendall Tau of BLEU, Meteor, and BLEURT with human judgements on the WebNLG dataset, varying the size of the data used for training and validation.

Outline

1. Contribution
2. Method
3. Experiments
4. Conclusion

Conclusion

1) Designing crowd sourcing experiments is an expensive and high-latency process

2) So, NLG researchers commonly employ “automatic evaluation metrics” (e.g., BLEU and ROUGE)

3) However, theses metrics rely on N-gram overlap which are only sensitive to lexical variation

4) In “IID assumption”, automatic evaluation is problematic because of domain drifts, and quality drifts: a model trained on ratings data from 2015 may fail to distinguish top performing systems in 2019

Thank you

https://jeiyoon.github.io/