# BLEURT

Notes by Trexd

# Overview

- BLEURT is a metric used to evaluate model based on the popular transformer model, BERT
- The authors claim that BLEURT is the new SOTA and can better model human judgements than other natural language metrics such as BLEU and ROUGE
- They accomplish this using synthetic data for pretraining stage.

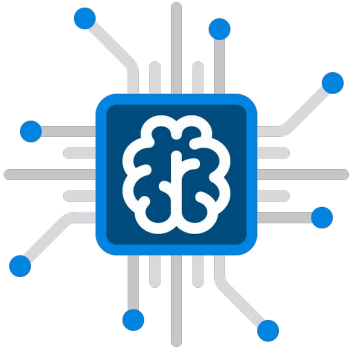
# Purpose

## Humans:

| Pros | Cons |
|---|---|
| • Accurate | • Expensive<br>• Time Consuming |

## Automated Metrics:

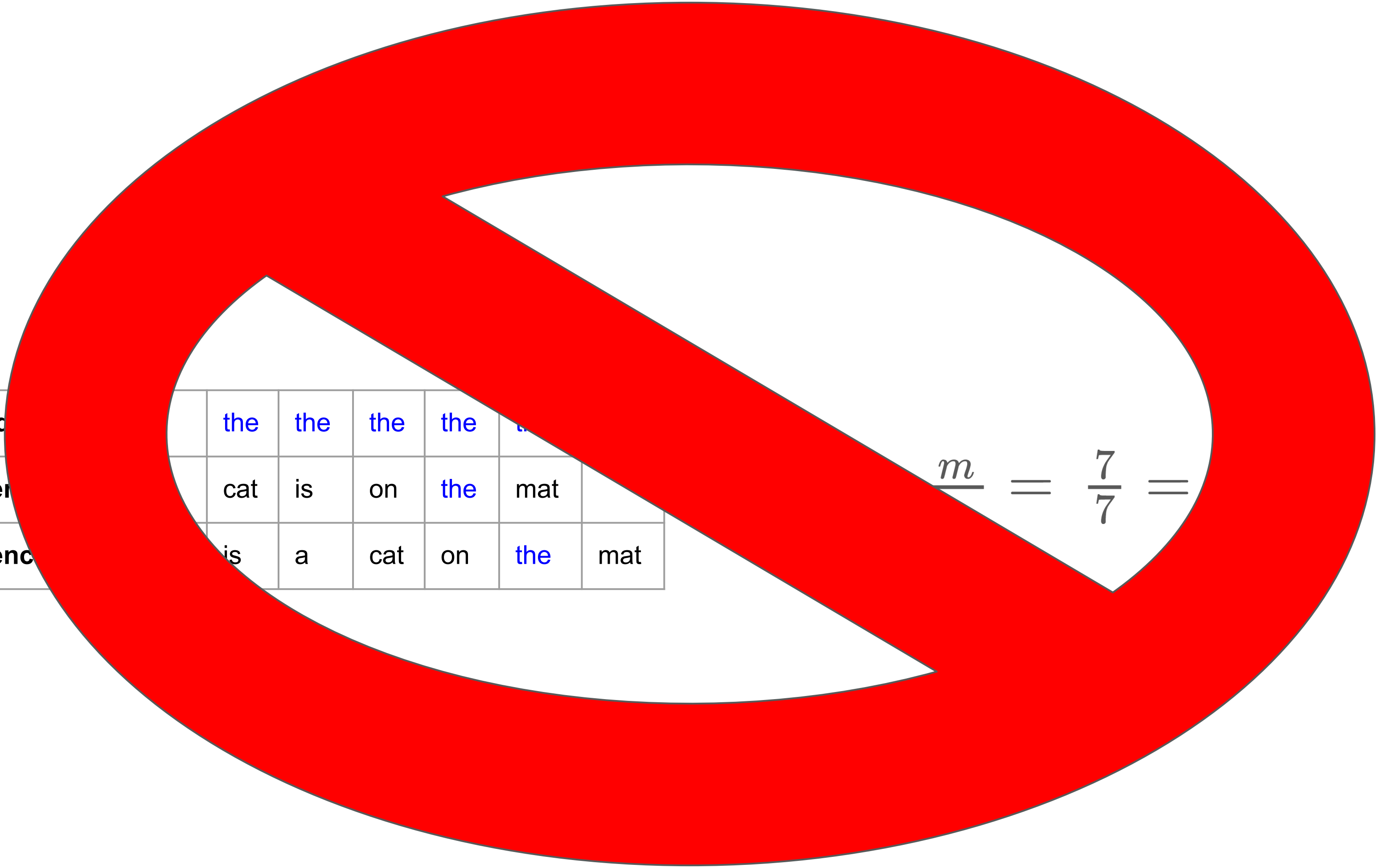
| Pros | Cons |
|---|---|
| • Fast<br>• Cheap | • Subpar Accuracy compared to humans |



# Previous Metrics

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Candidate** | the | the | the | the | the | the | the |
| **Reference 1** | the | cat | is | on | the | mat | |
| **Reference 2** | there | is | a | cat | on | the | mat |

$$\mathrm{P} = \frac{m}{w_t} = \frac{7}{7} = 1$$



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Candid** | [illegible] | the | the | the | the | th | [illegible] |
| **Referer** | [illegible] | cat | is | on | the | mat | |
| **Referenc** | [illegible] | is | a | cat | on | the | mat |

$$\frac{m}{} = \frac{7}{7} =$$

# BLEU (2002)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Candidate** | the | the | the | the | the | the | the |
| **Reference 1** | the | cat | is | on | the | mat | |

$$m_{max} = 2,\ m_w = 7$$

$$P = \frac{2}{7}$$

# Bleu score on bigrams

Example: Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

MT output: The cat the cat on the mat.

| | Count | $Count_{clip}$ |
|---|---|---|
| the cat | 2 | 1 |
| cat the | 1 | 0 |
| cat on | 1 | 1 |
| on the | 1 | 1 |
| the mat | 1 | 1 |

$$\frac{4}{6}$$

We compute the brevity penalty BP,

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Then,

$$\textsc{Bleu} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

# BLEURT (No “Priming”)

# BLEURT (No "Priming")

$x$ : reference
$\hat{x}$ : prediction

$y$ : human score

Reference: The cat is very happy

Prediction: 猫はとても幸せです

Human Score: 1.00

# BLEURT ("No Priming")

- Now we have a dataset...

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

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

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{N}\sum_{n=1}^{N} ||y_i - \hat{y}||^2$$

- …but it's too small for BERT fine-tuning

# Synthetic “Priming”

# Synthetic “Priming”

- Random perturbations on Wikipedia (1.8 million segments)
- Completed after pre-training but before fine-tuning
- Consists of:
  - 1. Mask-filling (15 masks per sentence)
    - Randomly mask a tokens in a given sentence
    - Mask sections of a sentence
  - 2. Backtranslation
    - Eg. English → French → English
  - 3. Randomly Dropping words

WIKIPEDIA
The Free Encyclopedia

# Pre-Training Signals

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

Table 1: Our pre-training signals.

# Backtranslation Likelihood

$P(\hat{z} \mid z)$ - Probability that z_hat is a backtranslation of z

$P_{\mathrm{en}\to\mathrm{fr}}(z_{\mathrm{fr}} \mid z)$

$P_{\mathrm{fr}\to\mathrm{en}}(z \mid z_{\mathrm{fr}})$ - Two Language Models

$$P(\hat{z} \mid z) = \sum_{z_{\mathrm{fr}}} P_{\mathrm{fr}\to\mathrm{en}}(\hat{z} \mid z_{\mathrm{fr}}) P_{\mathrm{en}\to\mathrm{fr}}(z_{\mathrm{fr}} \mid z)$$

$$z^*_{\mathrm{fr}} = \mathrm{argmax} P_{\mathrm{en}\to\mathrm{fr}}(z_{\mathrm{fr}} \mid z)$$

$$P(\tilde{z} \mid z) \approx P_{\mathrm{fr}\to\mathrm{en}}(\tilde{z} \mid z^*_{\mathrm{fr}}) \longrightarrow \tau_{\mathrm{en}\to\mathrm{fr}}, \tilde{z} \mid z = \frac{\log(\mathrm{P}(\tilde{z} \mid z))}{|\tilde{z}|}$$

# Textual Entailment

- Does statement B entail statement A?

| | |
|---|---|
| **Statement A** | “I will be 28 this year” |
| **Statement B** | “I am currently living” |

$$\tau_{\text{entail}} = \left(\tau_{\text{Entail}}, \tau_{\text{Contridict}}, \tau_{\text{Neutral}}\right)$$

# Backtranslation flag

Is a Boolean that indicates whether the perturbation was generated with backtranslation or with mask-filling.

# Modeling - Putting it all together

- MSE for regression tasks
- Multiclass cross-entropy for classification tasks

All together:

$$\ell_{\text{pre-training}} = \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \gamma_k \ell_k(\boldsymbol{\tau}_k^m, \hat{\boldsymbol{\tau}}_k^m)$$

# Experiments and Results

# Quick Note on Metrics - “DARR”, Kendall’s Tau

1. Get all translations for a given reference segment and enumerate all pairs
2. Discard all similar scores (less than 25 points away on a 100 point scale)
3. For each remaining pair, they then determine which translation is the best according both human judgment and the candidate metric

- |Concordant| = number of pairs where NLG metrics **agree**
- |Discordant| = number of pairs where NLG metrics **disagree**

$$\frac{|\text{Concordant}| - |\text{Discordant}|}{|\text{Concordant}| + |\text{Discordant}|}$$

# Models and Results

1. Bleurt: BERT-Large (24 layers, 1024 hidden units, 16 heads)
2. BLEURTbase: (12 layers, 768 hidden units, 12 heads)

- “Pre” means no priming

| model | de-en | fi-en | gu-en | kk-en | lt-en | ru-en | zh-en | **avg** |
|---|---|---|---|---|---|---|---|---|
| | $\tau$ / DA | $\tau$ / DA | $\tau$ / DA | $\tau$ / DA | $\tau$ / DA | $\tau$ / DA | $\tau$ / DA | $\tau$ / 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** |

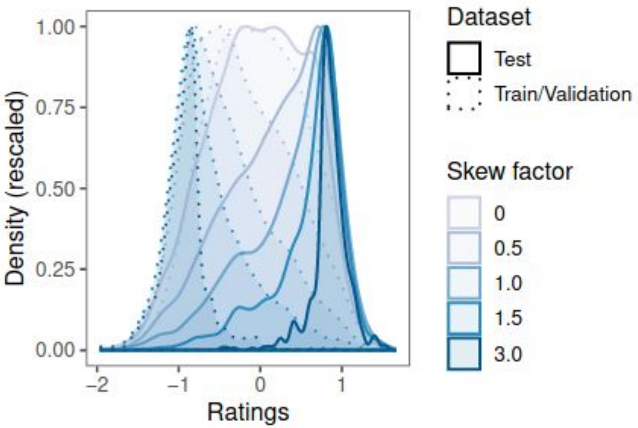



Figure 1: Distribution of the human ratings in the train/validation and test datasets for different skew factors.

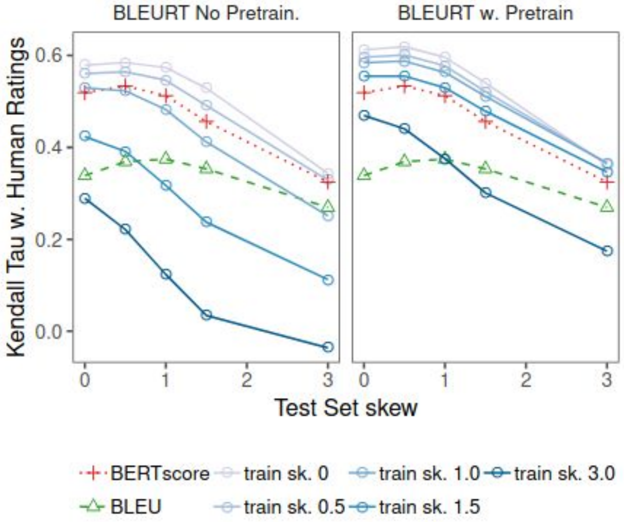



Figure 2: Agreement between BLEURT and human ratings for different skew factors in train and test.

# How skew works

- Sample the training and testing sets

1. Split the data into 10 bins of equal size
2. Sample using the following probabilities for the train and test sets

$$\frac{1}{B^a} \qquad \frac{1}{(11-B)^a}$$

Where B is the bin index and a is the skew factor

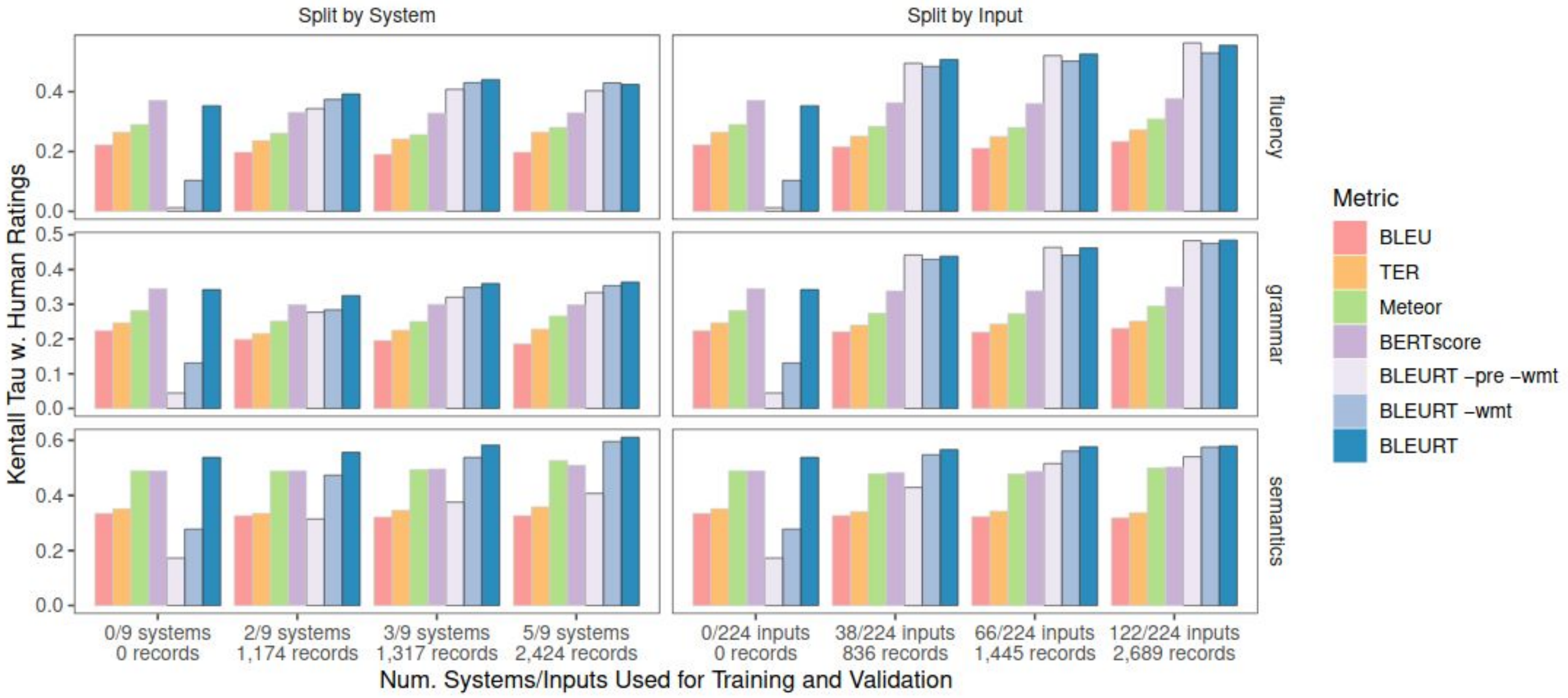



Figure 3: Absolute Kendall Tau of BLEU, Meteor, and BLEURT with human judgements on the WebNLG dataset, varying the size of the data used for training and validation.

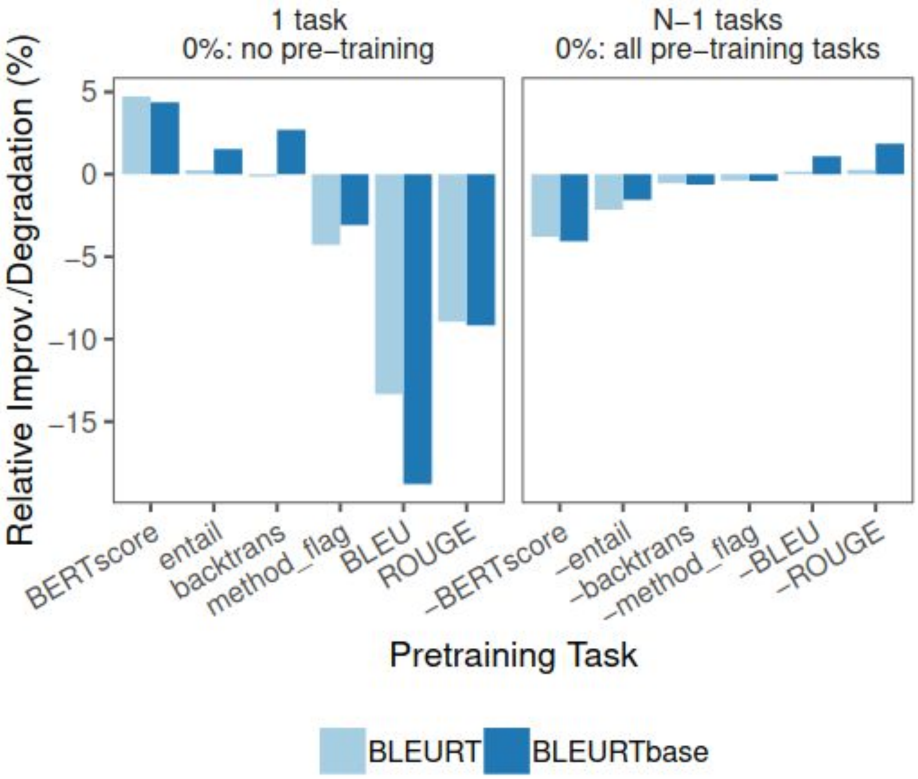



Figure 4: Improvement in Kendall Tau on WMT 17 varying the pre-training tasks.