# Evaluation @ EMNLP 2020 (Part 1)

# Outline

- BLEU might be Guilty but References are not Innocent. EMNLP 2020
- Unsupervised Reference-Free Summary Quality Evaluation via Contrastive Learning. EMNLP 2020
- Some thoughts on Evaluation Methods for Short-text Conversation.

# BLEU might be Guilty but References are not Innocent. EMNLP 2020.

*Markus Freitag, David Grangier, Isaac Caswell.*

Google Research

# Introduction

- Though choice of metric is important, the nature of the references is also critical.
- Typical references exhibit poor diversity, concentrating around translationese(翻译腔) language.
- This work studies how different reference collection methods impact the reliability of automatic evaluation.

# Contributions

- They collect different types of references on the same test set and show that it is possible to report strong correlation between automated evaluation with human metrics, even for high accuracy systems.
- They show that paraphrasing translations, when done carefully, can improve the quality of automated evaluations more broadly.
- They present an alternative multi-reference formulation that is more effective than multi reference BLEU for high quality output.

# Collecting High Quality and Diverse References

How we acquired additional references?

- **Increasing reference quality:** ask a professional translation service to provide an additional reference translation.
- **Diversified, natural references through paraphrasing:** use the same service to paraphrase existing references, asking a different set of linguists.

# Increasing reference quality

- A professional translation service was asked to create additional high quality references to measure the effect of different reference translations.
- The collection of additional references not only may yield better references, but also allows us to conduct various types of multi-reference evaluation.

# Diversified, natural references through paraphrasing

- They explore collecting diverse references using paraphrasing to steer away from translationese.
- To cover a wider diversity of target sentences, they first asked linguists to make only minor changes when paraphrasing sentences, and then changed the instructions to paraphrase the sentence as much as possible.

# Paraphrase the sentence as much as possible

## Instruction-1:

- 1. Read the sentence several times to fully understand the meaning
- 2. Note down key concepts
- 3. Write your version of the text without looking at the original
- 4. Compare your paraphrased text with the original and make minor adjustments to phrases that remain too similar

# Paraphrase the sentence as much as possible

**Instruction-2:**

- 1. Start your first sentence at a different point from that of the original source (if possible)
- 2. Use as many synonyms as possible
- 3. Change the sentence structure (if possible)

# Paraphrase the sentence as much as possible

| | |
|---|---|
| Source | The Bells of St. Martin’s Fall Silent as Churches in Harlem Struggle. |
| Translation | Die Glocken von St. Martin verstummen, da Kirchen in Harlem Probleme haben. |
| Paraphrase | Die Probleme in Harlems Kirchen lassen die Glocken von St. Martin verstummen. |
| Paraphrase | Die Kirchen in Harlem kampfen mit Problemen, und so lauten die Glocken von St. Martin nicht mehr. |

# Experimental Setup

- Data and Models
  - use the official submissions of the WMT 2019 English→German news translation task (Barrault et al., 2019) to measure automatic scores for different kinds of references.
- Human Evaluation
  - Human raters are asked to assess a given translation by how adequately it expresses the meaning of the corresponding source sentence on an absolute 0-100 rating scale.

# Experiments

- Three additional references are generated for the WMT 2019 English→German news translation task.
- In addition to acquiring an additional reference (AR), they also asked linguists to paraphrase the existing WMT reference (WMT.p) and the AR reference (AR.p).

# Human Evaluation of References

- The paraphrased references are rated as slightly less accurate. This may at least in part be an artifact of the rating methodology.
- The combined paraphrased reference HQ(P) has a higher human rating than WMT or AR alone.

| | adequacy rating |
|---|---|
| WMT | 85.3 |
| WMT.p | 81.8 |
| AR | 86.7 |
| AR.p | 80.8 |
| HQ(R) [WMT+AR] | 92.8 |
| HQ(P) [WMT.p+AR.p] | 89.1 |
| HQ(all 4) [all 4] | 95.3 |

Table 2: Human adequacy assessments for different kinds of references, over the full set of 1997 sentences. HQ(P) has been generated by picking sentence-by-sentence the more accurate rated translation from WMT.p and AR.p. HQ(R) and HQ(all 4) have been generated with the same method by either combining WMT and AR or all four reference translations.

# Correlation with Human Judgement

- All 3 new references (AR, WMT.p, AR.p) show higher correlation than the original WMT reference.
- Each paraphrased reference set shows higher correlation when compared to the “standard” reference set.

| Full Set (22) | Reference | $\rho$ | $\tau$ |
|---|---|---|---|
| single ref | WMT | 0.88 | 0.72 |
| | AR | 0.89 | 0.76 |
| | WMT.p | **0.91** | **0.79** |
| | AR.p | 0.89 | 0.77 |
| single ref | HQ(R) | **0.91** | 0.78 |
| | HQ(P) | **0.91** | 0.78 |
| | HQ(all 4) | **0.91** | **0.79** |
| multi ref | AR+WMT | 0.90 | 0.75 |
| | AR.p+WMT.p | 0.90 | **0.79** |
| | all 4 | 0.90 | 0.75 |

Table 3: Spearman’s $\rho$ and Kendall’s $\tau$ for the WMT2019 English→German official submissions with human ratings conducted by the WMT organizers.

# Correlation with Human Judgement

- Not one of the three combined references HQ(R), HQ(P), HQ(all 4) shows higher correlation than the paraphrased reference set WMT.p.
- If references are rated as more adequate, will such references yield more reliable automated scores ?

| Full Set (22) | Reference | $\rho$ | $\tau$ |
|---|---|---|---|
| single ref | WMT | 0.88 | 0.72 |
| | AR | 0.89 | 0.76 |
| | WMT.p | **0.91** | **0.79** |
| | AR.p | 0.89 | 0.77 |
| single ref | HQ(R) | **0.91** | 0.78 |
| | HQ(P) | **0.91** | 0.78 |
| | HQ(all 4) | **0.91** | **0.79** |
| multi ref | AR+WMT | 0.90 | 0.75 |
| | AR.p+WMT.p | 0.90 | **0.79** |
| | all 4 | 0.90 | 0.75 |

Table 3: Spearman's $\rho$ and Kendall's $\tau$ for the WMT2019 English→German official submissions with human ratings conducted by the WMT organizers.

# Correlation with Human Judgement

- Multi-reference BLEU does not exhibit better correlation with human judgments either than single-reference BLEU or than the composed reference sets HQ(x).

| Full Set (22) | Reference | $\rho$ | $\tau$ |
|---|---|---|---|
| single ref | WMT | 0.88 | 0.72 |
| | AR | 0.89 | 0.76 |
| | WMT.p | **0.91** | **0.79** |
| | AR.p | 0.89 | 0.77 |
| single ref | HQ(R) | **0.91** | 0.78 |
| | HQ(P) | **0.91** | 0.78 |
| | HQ(all 4) | **0.91** | **0.79** |
| multi ref | AR+WMT | 0.90 | 0.75 |
| | AR.p+WMT.p | 0.90 | **0.79** |
| | all 4 | 0.90 | 0.75 |

Table 3: Spearman's $\rho$ and Kendall's $\tau$ for the WMT2019 English→German official submissions with human ratings conducted by the WMT organizers.

# Alternative Metrics

- The paraphrased version of each reference set yields higher correlation with human evaluation across all evaluated metrics than the corresponding original references, with the only exception of TER for HQ(P).

| metric | WMT | HQ(R) | WMT.p | HQ(P) | HQ(all) |
|---|---|---|---|---|---|
| BLEU | 0.72 | 0.78 | **0.79** | **0.79** | **0.79** |
| 1 - TER | 0.71 | **0.74** | 0.71 | 0.67 | **0.74** |
| chrF | 0.74 | 0.81 | 0.78 | **0.82** | 0.78 |
| MET | 0.74 | **0.81** | **0.81** | **0.81** | 0.80 |
| BERTS | 0.78 | **0.82** | **0.82** | **0.82** | 0.81 |
| Yisi-1 | 0.78 | 0.84 | 0.84 | **0.86** | 0.84 |

Table 5: WMT 2019 English→German: Correlations (Kendall's $\tau$) of alternative metrics: BLEU, 1.0 - TER, chrF, METEOR, BERTScore, and Yisi-1.

# Why Paraphrases?

- All references generated with human translations (WMT, HQ(R) and HQ(all 4)) show negative correlation with human ratings.
- All references that rely purely on paraphrased references do produce the correct ranking of these three systems.

| Reference | bitext | APE | BT | correct? |
|---|---|---|---|---|
| human | 84.5 | 86.1 | 87.8 | ✓ |
| WMT | 39.4 | 34.6 | 37.9 | ✗ |
| WMT.p | 12.5 | 12.7 | 12.9 | ✓ |
| HQ(R) | 35.0 | 32.1 | 34.9 | ✗ |
| HQ(p) | 12.4 | 12.8 | 13.0 | ✓ |
| HQ(all 4) | 27.2 | 25.8 | 27.5 | ✗ |

Table 6: BLEU scores for WMT newstest 2019 English→German for MT systems trained on bitext, augmented with BT or using APE as text naturalizer. The *correct* column indicates if the model ranking agrees with human judgments.

# Conclusions

- The paraphrased references result in more reliable automated evaluations.
- The paraphrased references are able to correctly distinguish baselines from systems known to produce more natural output (those augmented with either BT or APE).
- Multi-reference BLEU does not exhibit better correlation with human judgments than single-reference BLEU.

# Unsupervised Reference-Free Summary Quality Evaluation via Contrastive Learning. EMNLP 2020

*Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, Shouling Ji*

Zhejiang University, IBM Research

# Introduction

- Previous approaches, such as ROUGE, mainly consider the informativeness of the assessed summary and require human-generated references for each test summary.
- This work proposes a new metric to evaluate the summary qualities without reference summaries by unsupervised contrastive learning.
- To learn the metric, for each summary, different types of negative samples with respect to different aspects of the summary qualities are constructed and the model is trained with a ranking loss.

# Model Framework

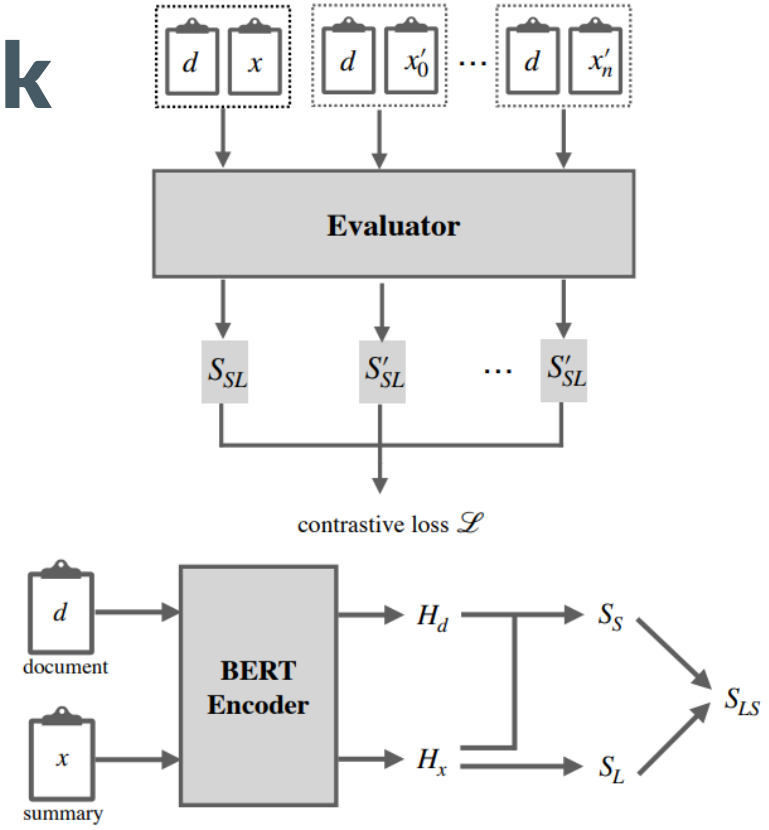


Figure 1: Model Framework. The top figure describes the framework for contrastive learning, where for each document $x$, we create different types of negative samples and compare them with $x$ to get a ranking loss. The bottom figure is the evaluator which generates the final evaluation score. For short, here we use $S_S$, $S_L$ and $S_{LS}$ to indicate $S\_Score$, $L\_Score$ and $LS\_Score$.

# Evaluating Semantic Quality

Formally, let $S_x$ and $S_d$ be the sequence of tokens in the summary $x$, and the source document $d$. A sequence of tokens is encoded into a sequence of token embeddings $H$ by the BERT encoder:

$$H_x = BERT(S_x)$$

$$H_d = BERT(S_d)$$

# Evaluating Semantic Quality

The semantic quality of the target summary is measured by calculating the semantic similarity between $x$ and its source document $d$:

$$S\_Score(x) = Sim(H_d^0, H_x^0)$$

# Evaluating linguistic quality

First use the BERT encoder to get the representation of the summary:

$$H_x = BERT(x)$$

Then calculate the probability of the sequence based on this representation:

$$P_x = softmax(W_1^T(\sigma(W_0^T H_x)))$$

# Evaluating linguistic quality

Motivated by the perplexity, the linguistic quality of $x$ can be calculated as:

$$L\_Score(x) = \frac{1}{|x|} \sum_{i}^{n} logp_x^i$$

# Evaluating Both Dimension

The final metric is developed by linearly combining the $S\_Score$ and $L\_Score$:

$$LS\_Score(x) = \alpha L\_Score(x) + \beta S\_Score(x)$$

In this work, they fix $\alpha = 0.01$ and $\beta = 1$ to scale the $L\_Score$ and the $S\_Score$

# Contrastive Training

- A new unsupervised training framework via contrastive learning is developed to alleviate the requirement of reference summaries as well as given human evaluation scores.
- For a given good summary, it is easy to create a summary with worse quality ( e.g. disordering the words/sentences).
- Then we can compare these two summaries to get a contrastive loss:

$$Loss = \sum_{r \in R} \sum_{\hat{x} \in \hat{X}_r} max(0, 1 - (LS\_Score(r) - LS\_Score(\hat{x})))$$

# Contrastive Training

| |
|---|
| **Original summary:** |
| Kristina Patrick from Alaska filmed her German Shepherd Pakak performing a very skillful trick. Footage shows the pup taking the ball from her mouth with her paws and holding it up high in the air to admire it. She then carefully lowers it back down to the starting point. |
| **Negative samples:** |
| 1. delete words |
| Patrick ∧ from Alaska filmed her German Shepherd Pakak performing a very skillful trick. Footage shows the pup taking the ∧ from her ∧ with her paws and holding it up high in the air to ∧ it. She then carefully lowers it back down to the starting point. |
| 2. add sentences |
| Kristina Patrick from Alaska filmed her German Shepherd Pakak performing a very skillful trick. Footage shows the pup taking the ball from her mouth with her paws and holding it up high in the air to admire it. She then carefully lowers it back down to the starting point. ~~PAKAK 's owner says she loves playing with balls.~~ |
| 3. disorder words |
| Kristina Patrick skillful Alaska filmed her performing Shepherd a German Pakak very from trick. Footage shows the pup taking the ball from admire mouth with and paws her holding it up high her to air the in it. She then back lowers it carefully to down the starting point. |

Table 2: An example of negative sampling.

# Questions

- Does the contrastive learning method obtain better performance over other baselines even without reference summaries?
- Can the evaluator capture the expected aspects of summary qualities, and does it outperform others under the same contrastive learning framework?
- Is the method generalizable to different datasets? That is, how does it perform if we train the metric on one dataset and test on another one?

# Experimental Settings

- Datasets
  - Newsroom
  - CNN/Daily Mail
- Baselines
  ROUGE, METEOR, BERTScore, WMS/SMS/S+WMS, MoverScore, BERT+Cos+Ref, BERT+Cos+Doc

# Results on Newsroom

- $LS\_Score$ achieves best correlations in all of the different dimensions.

| | Coh. | Flu. | Inf. | Rel. |
|---|---|---|---|---|
| **ROUGE-1** | 0.2446 | 0.1991 | 0.3371 | 0.3028 |
| **ROUGE-2** | 0.1133 | 0.0763 | 0.1816 | 0.1385 |
| **ROUGE-L** | 0.2164 | 0.1736 | 0.3178 | 0.2700 |
| **METEOR** | 0.3325 | 0.3347 | 0.4424 | 0.4117 |
| **BERTScore-R** | 0.2355 | 0.2227 | 0.2972 | 0.2787 |
| **BERTScore-P** | -0.0263 | -0.0221 | -0.0215 | -0.0302 |
| **BERTScore-F** | 0.1206 | 0.1072 | 0.1681 | 0.1426 |
| **WMS** | 0.2389 | 0.2355 | 0.3003 | 0.2406 |
| **SMS** | 0.2394 | 0.2400 | 0.2946 | 0.2401 |
| **S+WMS** | 0.2433 | 0.2405 | 0.3022 | 0.2432 |
| **MoverScore** | 0.1458 | 0.1021 | 0.2070 | 0.1724 |
| **BERT+Cos+Ref** | 0.0452 | 0.0333 | 0.0475 | 0.0534 |
| **BERT+Cos+Doc** | 0.3998 | 0.3492 | 0.4530 | 0.4279 |
| **LS_Score** | **0.6390** | **0.5933** | **0.7163** | **0.6563** |

Table 4: Spearman correlation w.r.t. coherence (Coh.), fluency (Flu.), informativeness (Inf.) and relevancy (Rel.) on Newsroom. Best results are in bold.

# Results on CNN/Daily Mail

- $LS\_Score$ still achieves best correlations in all of the different dimensions.

| | Overall | Grammar | Redundancy |
|---|---|---|---|
| **ROUGE-1** | 0.1953 | 0.0975 | 0.2174 |
| **ROUGE-2** | 0.1355 | 0.0701 | 0.1442 |
| **ROUGE-L** | 0.1925 | 0.0973 | 0.2072 |
| **METEOR** | 0.0773 | 0.0173 | 0.1147 |
| **BERTScore-R** | 0.2628 | 0.1721 | 0.2780 |
| **BERTScore-P** | 0.1754 | 0.1828 | 0.1180 |
| **BERTScore-F** | 0.2536 | 0.2041 | 0.2348 |
| **WMS** | 0.1809 | 0.1080 | 0.2274 |
| **SMS** | 0.1814 | 0.1021 | 0.2313 |
| **S+WMS** | 0.1830 | 0.1075 | 0.2314 |
| **MoverScore** | 0.2220 | 0.1522 | 0.2289 |
| **BERT+Cos+Doc** | 0.1484 | 0.1110 | 0.1237 |
| **BERT+Cos+Ref** | 0.2130 | 0.1316 | 0.2284 |
| **LS_Score** | **0.3342** | **0.2664** | **0.2875** |

Table 5: Spearman correlation on CNN/Daily Mail.

# Ablation Study for Evaluator Selection

- BERT+Linear uses a linear regressor to map the BERT embeddings of summaries into a score. The model is trained under the same contrastive learning framework.
- The proposed model is superior to BERT+Linear a lot in most cases.

| | Coh. | Flu. | Inf. | Rel. |
|---|---|---|---|---|
| **Bert+Linear** | 0.4213 | 0.4511 | 0.3075 | 0.3400 |
| **LS_Score** | **0.6390** | **0.5933** | **0.7163** | **0.6563** |

Table 6: Ablation studies on Newsroom. The models use the same contrastive learning framework but different evaluators.

| | Overall | Grammar | Redundancy |
|---|---|---|---|
| **Bert+Linear** | 0.2711 | **0.2886** | 0.1664 |
| **LS_Score** | **0.3342** | 0.2664 | **0.2875** |

Table 7: Ablation studies on CNN/Daily Mail. The models use the same contrastive learning framework but different evaluators.

# Cross-dataset Transferability

- The cross-data training makes the performance of $LS\_Score$ cross slightly lower than the original $LS\_Score$ in most cases, but it still outperform all other baselines.

| | Coh. | Flu. | Inf. | Rel. |
|---|---|---|---|---|
| **ROUGE-1** | 0.2446 | 0.1991 | 0.3371 | 0.3028 |
| **ROUGE-L** | 0.2164 | 0.1736 | 0.3178 | 0.2700 |
| **BERTScore-R** | 0.2355 | 0.2227 | 0.2972 | 0.2787 |
| **MoverScore** | 0.1458 | 0.1021 | 0.2070 | 0.1724 |
| **BERT+Cos+Doc** | 0.3998 | 0.3492 | 0.4530 | 0.4279 |
| **LS_Score** | **0.6390** | **0.5933** | **0.7163** | **0.6563** |
| **LS_Score_cross** | *0.6271* | *0.5852* | *0.7008* | *0.6381* |

Table 8: Cross-dataset training results: Spearman correlation on Newsroom. The model of LS_Score_cross is trained on CNN/Daily Mail.

# Cross-dataset Transferability

- The cross-data training makes the performance of $LS\_Score$ cross slightly lower than the original $LS\_Score$ in most cases, but it still outperform all other baselines.

| | Overall | Grammar | Redundancy |
|---|---|---|---|
| **ROUGE-1** | 0.1953 | 0.0975 | 0.2174 |
| **ROUGE-L** | 0.1925 | 0.0973 | 0.2072 |
| **BERTScore-R** | 0.2628 | 0.1721 | 0.2780 |
| **MoverScore** | 0.2220 | 0.1522 | 0.2289 |
| **BERT+Cos+Doc** | 0.1484 | 0.1110 | 0.1237 |
| **LS_Score** | **0.3342** | **0.2664** | 0.2875 |
| **LS_Score_cross** | *0.2874* | *0.1915* | ***0.2881*** |

Table 9: Cross-dataset training results: Spearman correlation on CNN/Daily Mail. The model LS_Score_cross is trained on Newsroom.

# Some thoughts on Evaluation Methods for Short-text Conversation

- It is generally assumed that metrics that support multiple references yield higher correlation with human judgements due to the increased diversity in the reference responses.
- Designing a metric that correlates well with human judgments for short-text conversation is very difficult, while constructing a diversified reference set is easier.

# Model Overview

- We can improve the performance of referenced metrics by enhancing the reliability of the reference set.
- Given a query sentence $\boldsymbol{x} = \{x_1, \ldots, x_m\}$, and a reference set $\{\hat{\boldsymbol{y}}_i\}_{i=1}^{N}$, our goal is to learn a function $\boldsymbol{f} : (\boldsymbol{x}, \{\hat{\boldsymbol{y}}\}) \rightarrow c$ that predicts a confidence score $c$ for each reference set.
- We simply generate a bunch of responses with existing response generation models, and score the responses with human judgment and the referenced metric repectively. The confidence scores for training data are correlation coefficients (Kendall Tau ($\tau$)/Pearson ( $r$)) between referenced metric and human ratings.

# Confidence Score Prediction

- We construct a graph with the query and the reference responses as nodes to consider the intrinsic variance between reference responses. Any two nodes with high cos-similarity will be connected by an undirected edge and the query node is connected to all the reference nodes.
- A GCN network coupled with a graph pooling layer is used to obtain the final representation of the graph. The final representation is then used to predict a confidence score $c$ for each reference set with a softmax classifier.

# How to use the Confidence Score?

- If a large nunmber of references are available, the quality of the reference set can be enhanced by expanding more references. The confidence score can help you check if the set is reliable enough.
- If large scale references are not available, we can collect responses that provides a big boost to the confidence score with retrieval methods. The responses can serve as templates for humans to rewrite, producing more appropriate responses.