Reference-Based Metrics Are Biased Against Blind and Low-Vision Users' Image Description Preferences

NLP4PI @ EMNLP 2024

Rhea Kapur
Department of Computer Science
Stanford University

Elisa Kreiss
Department of Communication
University of California, Los Angeles

4 people say…

1. A couple stands on the bank of a lake in Ludhiana, India.
2. A man and a woman stand at the edge of a lake in Ludhiana, India. The man is taking a photo of the woman, who is looking down and not at the camera. There are rocks in the bottom right corner of the image which are much closer to the camera than the people. Trees are reflected in the lake and there is a glimmer of sun, but overall it's muggy with partly cloudy skies.
3. A couple stands at the edge of a still water pond. The skies are blue with clouds and there are some stones not too far away from them. On the other side of the water body, there are a lot of trees.
4. Two people stand at the edge of a large lake surrounded by trees. The male-presenting figure takes a photo of the female-presenting person, where she has her head pointed down in a bashful manner and her hands are clasped behind her back. They seem to be a couple. There is a large pile of rocks protruding from the bottom right corner.

4 people say…

1. A couple stands on the **bank of a lake** in Ludhiana, India.
2. A man and a woman stand at the edge of a lake in Ludhiana, India. The man is taking a photo of the woman, who is looking down and not at the camera. There are rocks in the bottom right corner of the image which are much closer **to the camera** than the people. Trees are reflected in the lake and there is a glimmer of sun, but overall it's muggy with partly cloudy skies.
3. A couple stands at the edge of a still water pond. The skies are blue with clouds and there are some stones not too far away from them. On the other side of the water body, there are a lot of trees.
4. Two people stand at the edge of a large lake surrounded by trees. The male-presenting figure takes a photo of the female-presenting person, where she has her head pointed down in a bashful manner and her hands are clasped behind her back. They seem to be a couple. There is a large pile of rocks protruding from the bottom right corner.

4 ***reference descriptions*** (no image) say…

1. A couple stands on the bank of a lake in Ludhiana, India.
2. A man and a woman stand at the edge of a lake in Ludhiana, India. The man is taking a photo of the woman, who is looking down and not at the camera. There are rocks in the bottom right corner of the image which are much closer to the camera than the people. Trees are reflected in the lake and there is a glimmer of sun, but overall it's muggy with partly cloudy skies.
3. A couple stands at the edge of a still water pond. The skies are blue with clouds and there are some stones not too far away from them. On the other side of the water body, there are a lot of trees.
4. Two people stand at the edge of a large lake surrounded by trees. The male-presenting figure takes a photo of the female-presenting person, where she has her head pointed down in a bashful manner and her hands are clasped behind her back. They seem to be a couple. There is a large pile of rocks protruding from the bottom right corner.

New ***hypothesis*** description (no image)

Hypothesis: A man and a woman stand at the edge of a lake in front of some rocks. The man is taking a photo of the woman's face up close and from a high-up angle. The lake is surrounded by trees and greenery and the skies are partly cloudy.

Task: is the hypothesis a good description?

In reference-based metrics, we are solving a **similarity-measurement problem**: how similar is this hypothesis description to the four reference descriptions?

Reference-Based Metrics: Similarity Measurement

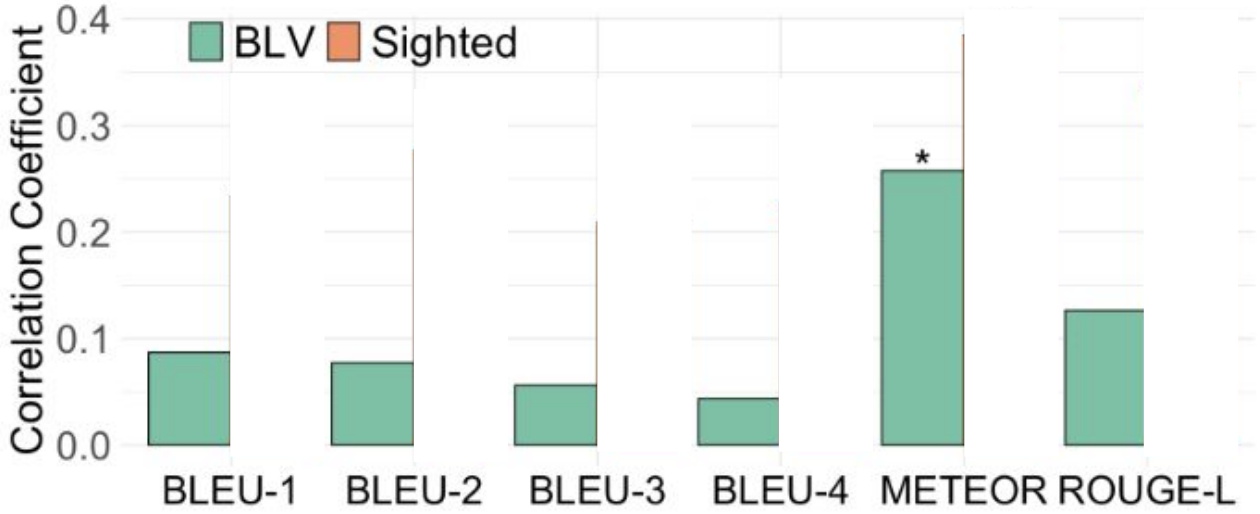
Reference-based metrics (BLEU, METEOR, ROUGE) make choices on **length, the number of references, and how to compute semantic similarity** to compute similarity:

- n-gram overlap
- synonym/morphological variant substitution
- brevity penalty

To what extent are the metrics able to differentiate between good and bad descriptions *for accessibility*? Do they reflect the preferences of blind and low vision (BLV) users?

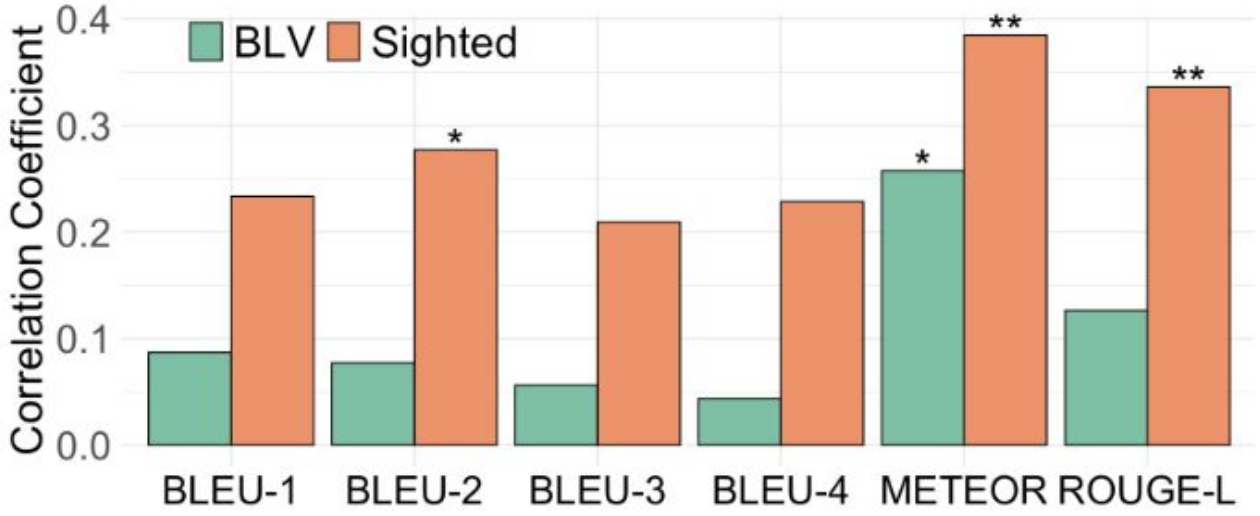
Dataset Creation and Correlations

- Building on Kreiss et al. (2022)
- Three datasets: context-sensitive, context-insensitive, context-sensitive with reference count varying exhaustively
- Correlations between BLV and sighted user ratings and BLEU, METEOR, ROUGE metrics
- Deep dive into hypothesis description length



Takeaway: Reference based metrics are weakly or not at all correlated with BLV user ratings.

*context-sensitive condition

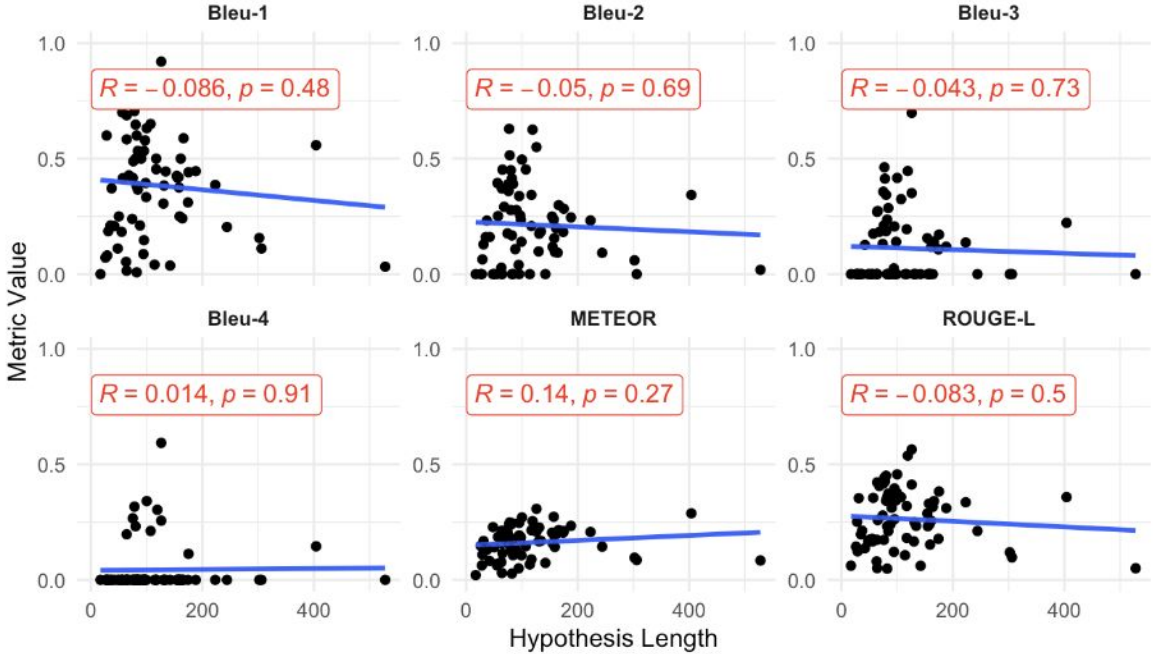


Takeaway: Reference based metrics are more often strongly and significantly correlated with sighted user ratings, and conversely are weakly or not at all correlated with BLV user ratings.

*context-sensitive condition

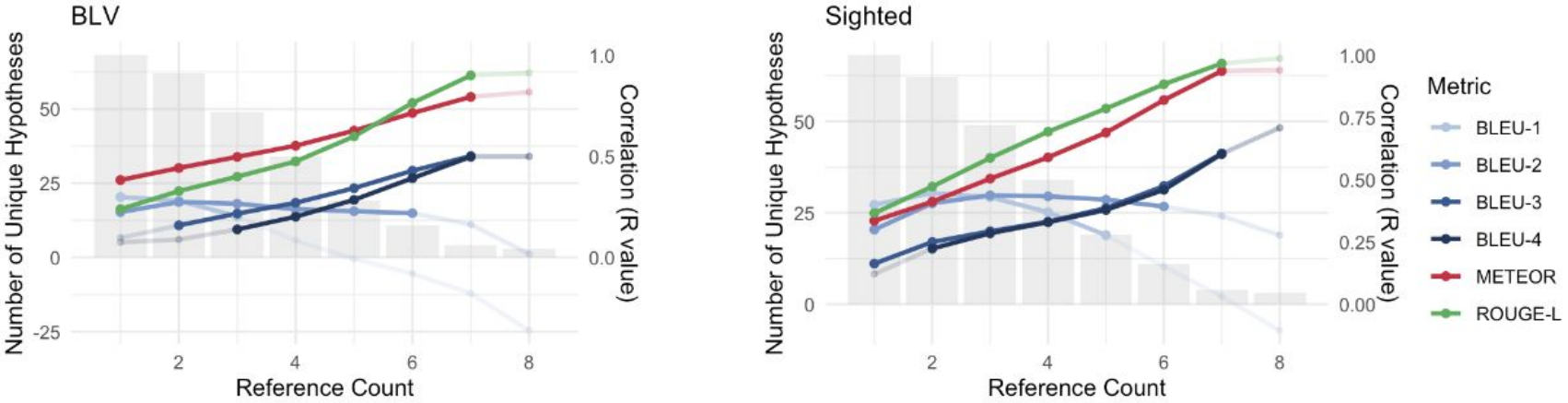
Takeaway: Context sensitivity varies greatly across metrics. Notably, there is divergent behavior with BLEU-3 and BLEU-4.

*context-**in**sensitive condition



Takeaway: No metric has strong correlations with hypothesis length, a factor strongly considered by BLV users.

*context-sensitive condition



Takeaway: There is a high degree of variation in how sensitive metrics are to reference count. Again, we see divergent behavior within BLEU and with BLEU-3/4 specifically.

*expanded context-sensitive condition

Future Paths

- Reference-based metrics are significantly biased toward sighted and against BLV user preferences → **Develop reference-based metrics which center BLV user in design and evaluation**
- More broadly, there are larger implications on **information density** (as informed by length and level of detail/informativity), who defines it, and who it serves

Thank you!

Code and data are available at https://github.com/rkapur102/reference-based-bias