Summary, The Moral of the Story

Metrics drive research:

- Measure the property that "saves money" in your application.
- Design automatic metrics to correlate with humans.

Comparisons of automatic scores trustworthy
only under all the following:

- a single test set was used (of your domain of interest),
- evaluated by a single evaluation tool (hopefully without bugs),
  E.g. for BLEU different tools tokenize and define ref. length differently.
- the metric reflects your final objective (AER vs. BLEU),
- confidence intervals are estimated.