Supplemental Material: How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation

Anonymous EMNLP submission

κ	# pairs	% pairs
> 0.2	253/253	100%
> 0.3	251/253	99.2%
> 0.4	225/253	88.9%
> 0.5	162/253	64.0%
> 0.6	50/253	19.8%
> 0.7	3/253	1.2%
> 0.8	0/253	0%

Table 1: Distribution of pairwise κ scores between each pair of human annotators, other than the annotators that were discarded due to low scores.

1 Distribution of kappa scores

Table 1 shows the full distribution over κ scores for each pair of human annotators. It is apparent that most of the scores (88.9%) are over 0.4, indicating a moderate agreement. This suggests that the task was reasonable and well understood by the annotators.

2 Full scatter plots

We present the scatterplots for all of the metrics consider and their correlation with human judgement, in Figures 3-7 below. As previously emphasized, there is very little correlation for any of the metrics, and the BLEU-3 and BLEU-4 scores are often close to zero.

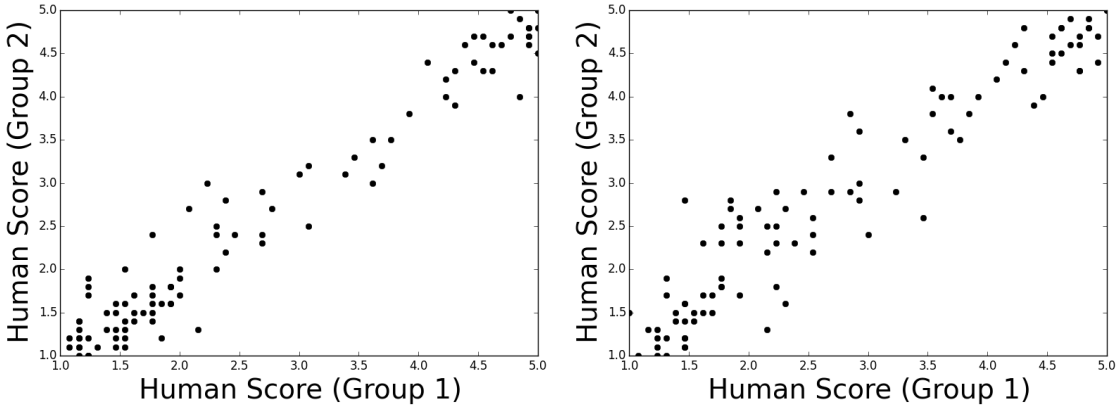
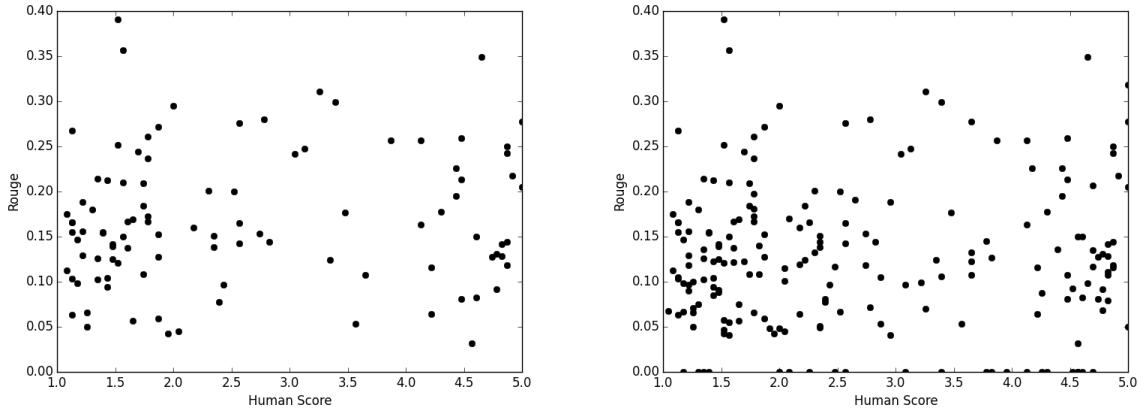
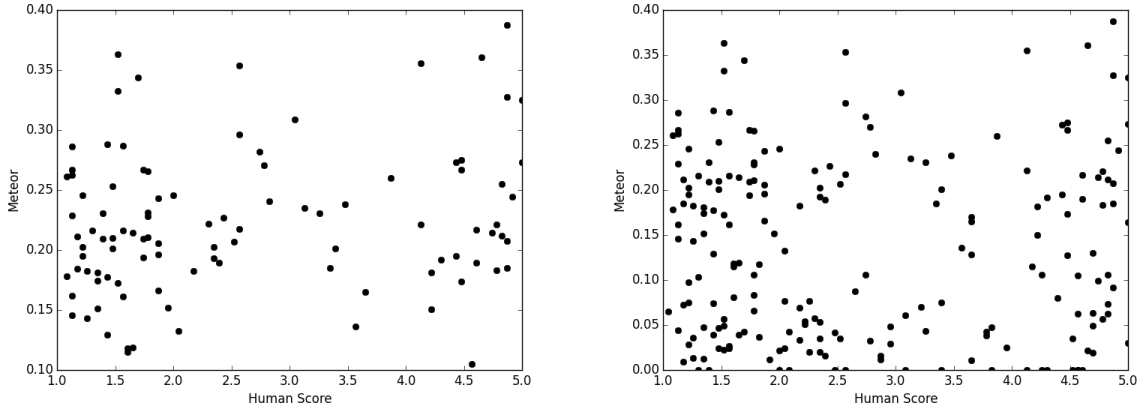


Figure 1: Scatter plots showing the correlation between two randomly chosen groups of human volunteers on the Twitter corpus (left) and Ubuntu Dialogue Corpus (right).

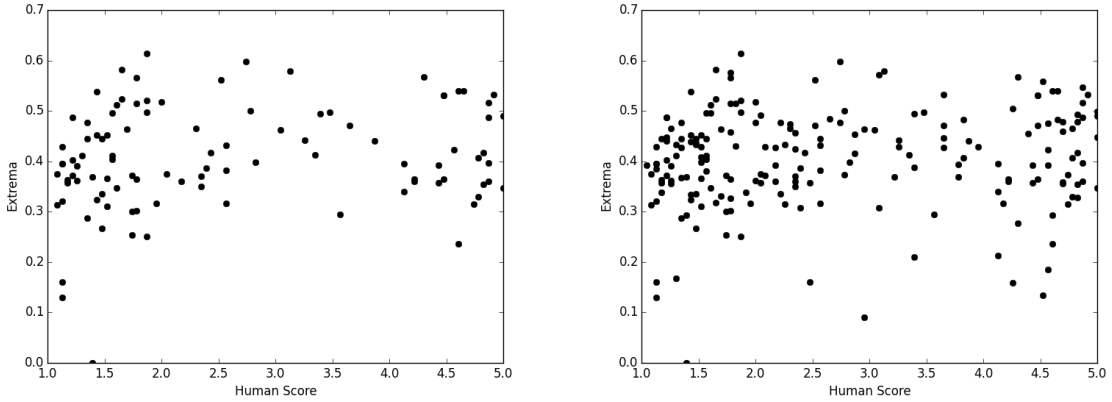


(a) ROUGE

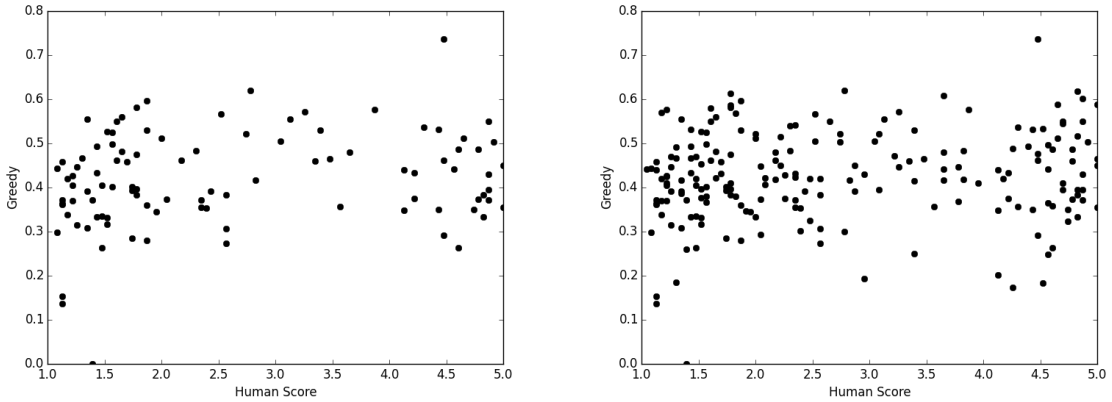


(b) METEOR

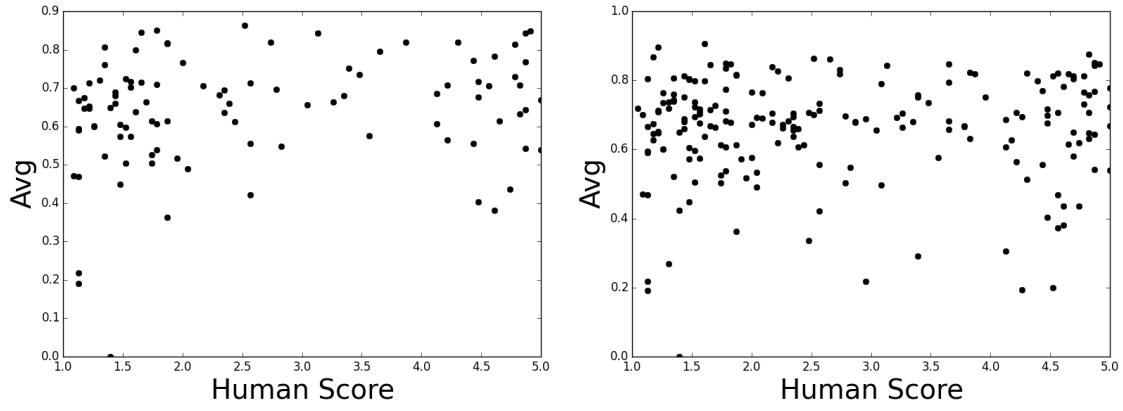
Figure 2: Scatter plots showing the correlation between metrics and human judgement on the Twitter corpus (left) and Ubuntu Dialogue Corpus (right). The plots represent ROUGE (a) and METEOR (b).



(a) Vector Extrema

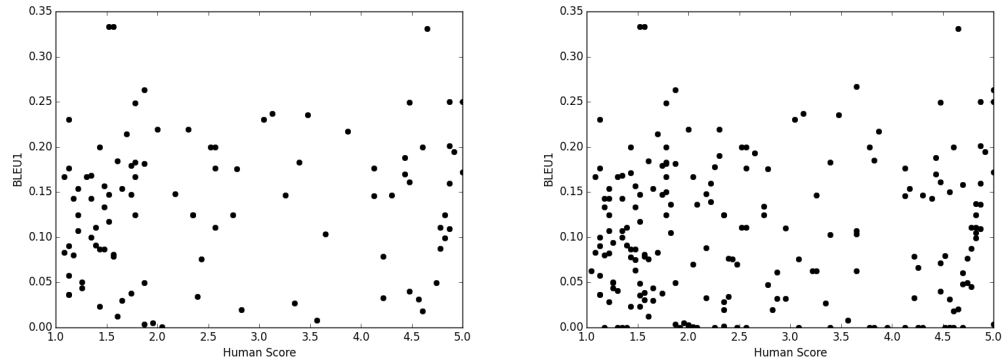


(b) Greedy Matching

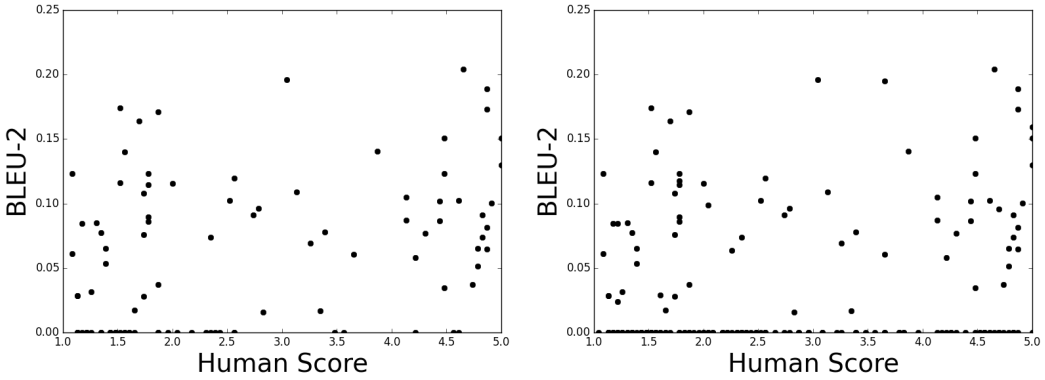


(c) Vector Averaging

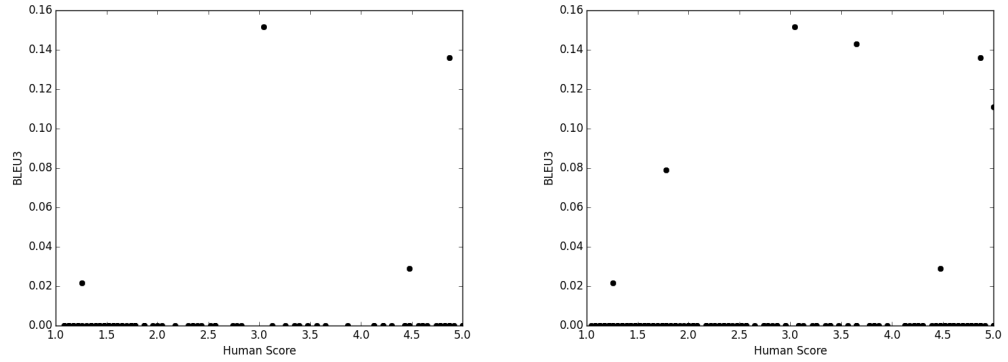
Figure 3: Scatter plots showing the correlation between metrics and human judgement on the Twitter corpus (left) and Ubuntu Dialogue Corpus (right). The plots represent vector extrema (a), greedy matching (b), and vector averaging (c).



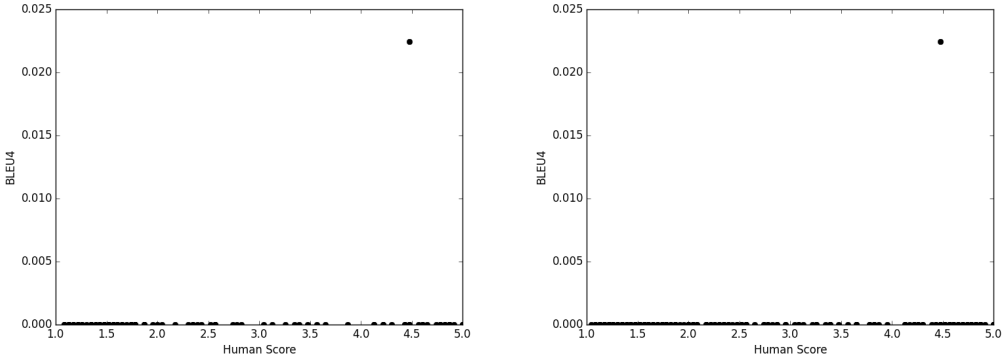
(a) BLEU-1



(b) BLEU-2



(c) BLEU-3



(d) BLEU-4

Figure 4: Scatter plots showing the correlation between metrics and human judgement on the Twitter corpus (left) and Ubuntu Dialogue Corpus (right). The plots represent BLEU-1 (a), BLEU-2 (b), BLEU-3 (c), and BLEU-4 (d).