

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.

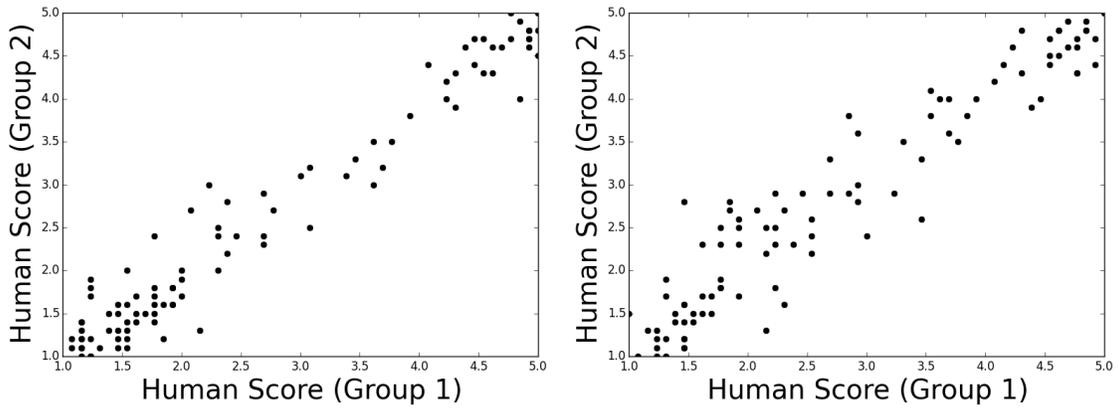
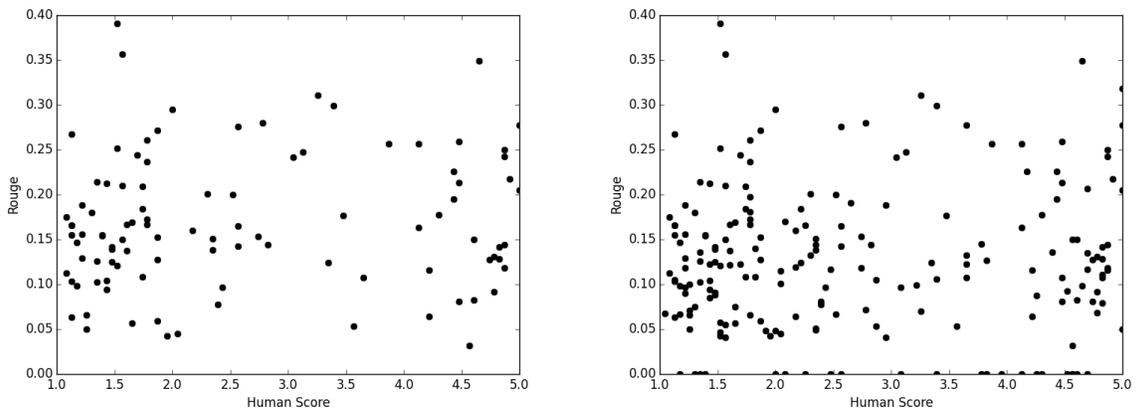
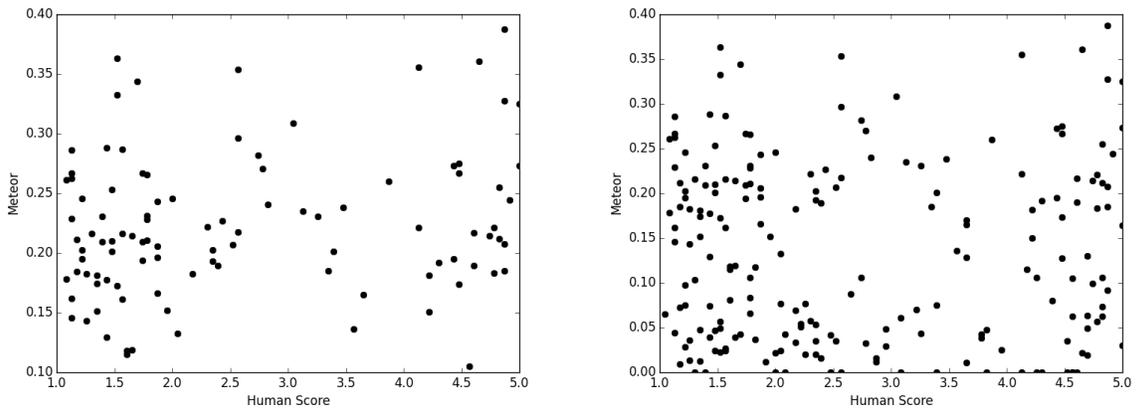


Figure 1: Scatter plots showing the correlation between two randomly chosen groups of human volunteers on the Twitter corpus (left) and Ubuntu Dialogue Corpus (right).



(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).

192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239

240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287

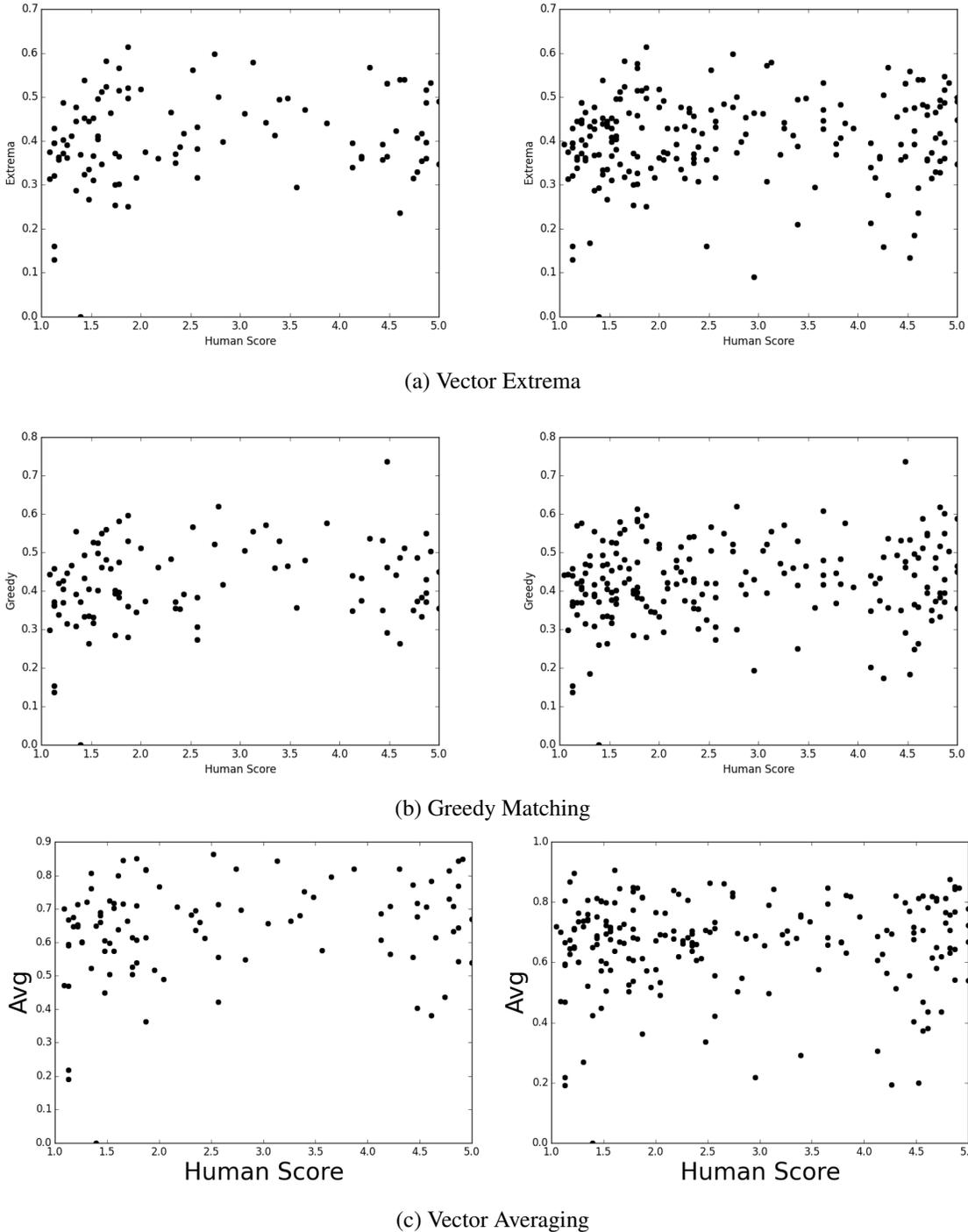
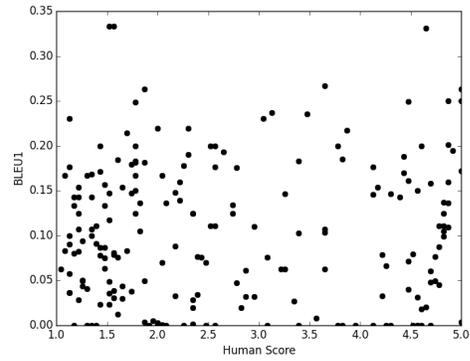
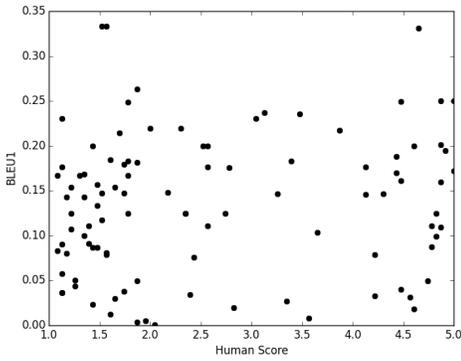


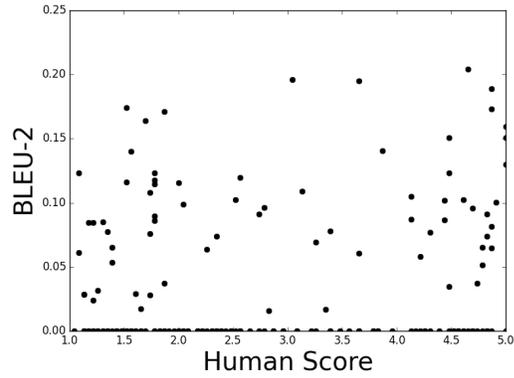
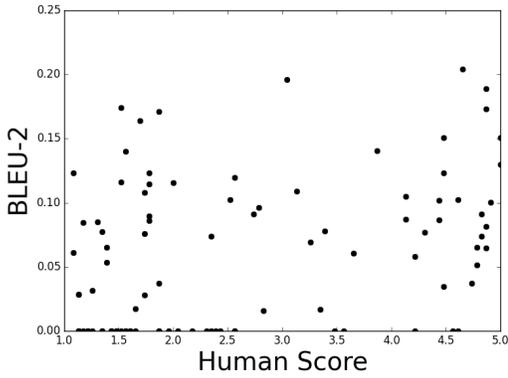
Figure 3: Scatter plots showing the correlation between metrics and human judgement on the Twitter corpus (left) and Ubuntu Dialogue Corpus (right). The plots represent vector extrema (a), greedy matching (b), and vector averaging (c).

288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335

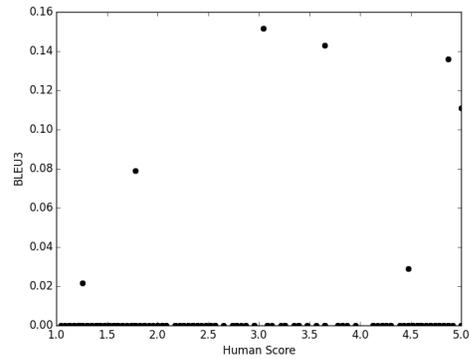
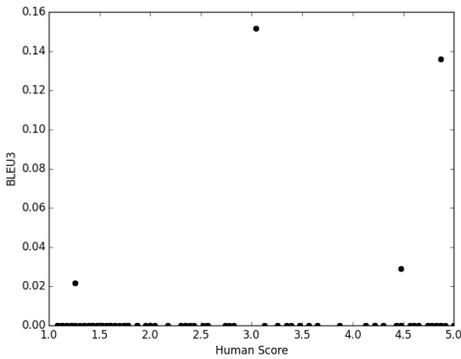
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383



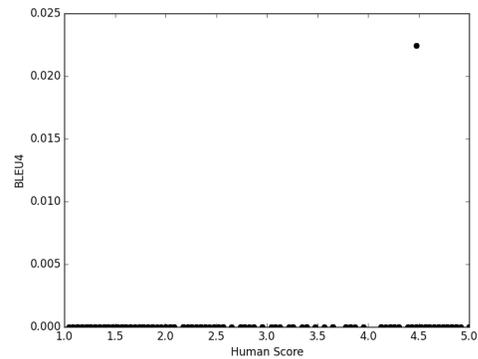
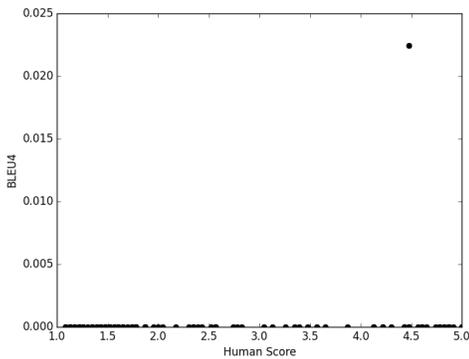
(a) BLEU-1



(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).