How Not To Evaluate Your Dialogue System

Ryan Lowe*, Iulian Serban*, Chia-Wei Liu*, Mike Noseworthy*,
Laurent Charlin, Joelle Pineau

McGill University & Université de Montréal

Evaluating Dialogue Systems

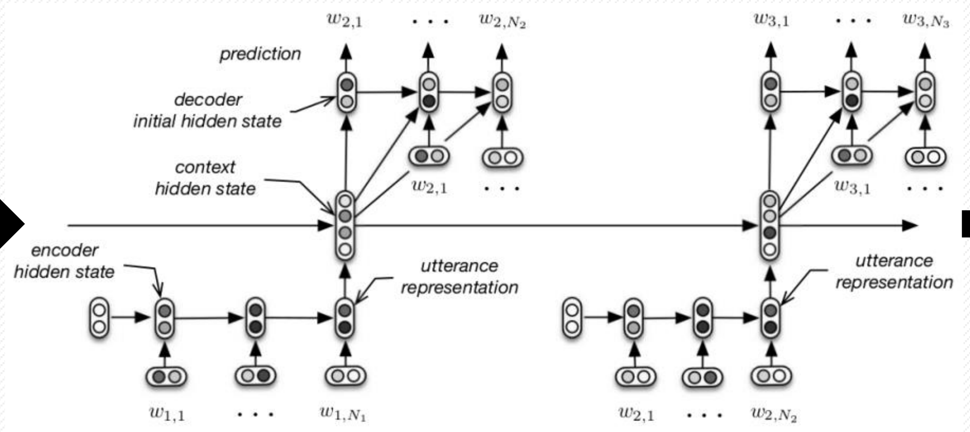
Focus on 'unsupervised' methods of evaluation, i.e. that do not require a supervised task completion signal:

- Human judgement
- Slot filling
- Retrieval (e.g. next utterance classification)
- Word perplexity
- Ground-truth utterance comparison

Comparison of ground-truth utterance

Context

Hey, want to go to the movies tonight?



Generated Response

Yeah, let's go see that movie about Turing!

Ground-truth response

Nah, I'd rather stay at home, thanks.

SCORE

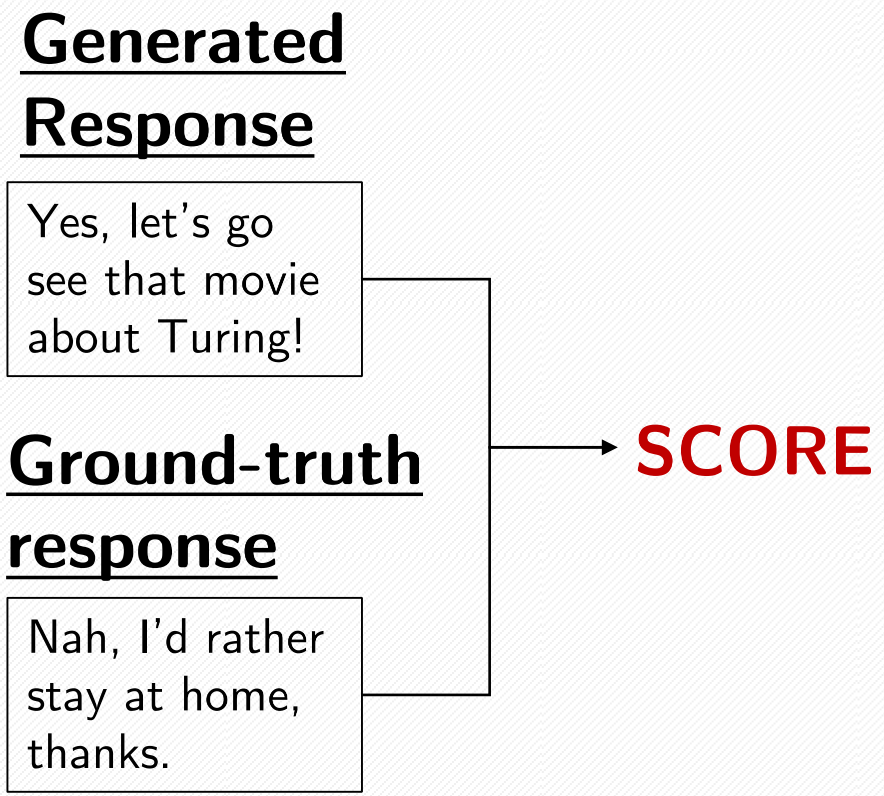
Comparison of ground-truth utterance

1) Word-overlap metrics:

- BLEU, METEOR, ROUGE

2) Word embedding-based metrics:

- Vector extrema, greedy matching, embedding average



Vector-based metrics

Assign score using (word2vec) embeddings of generated and ground-truth response:

1) Embedding average: compute sentence-level embeddings by taking average embedding + CD
2) Vector extrema: compute sentence-level embeddings by taking the extreme value of each dimension + CD
3) Greedy matching: greedily match word embeddings from each response (based on CD), take average score

Initial results

	Ubuntu Dialogue Corpus			Twitter Corpus		
	Embedding Averaging	Greedy Matching	Vector Extrema	Embedding Averaging	Greedy Matching	Vector Extrema
R-TFIDF	0.536 ± 0.003	0.370 ± 0.002	0.342 ± 0.002	0.483 ± 0.002	0.356 ± 0.001	0.340 ± 0.001
C-TFIDF	0.571 ± 0.003	0.373 ± 0.002	0.353 ± 0.002	0.531 ± 0.002	0.362 ± 0.001	0.353 ± 0.001
DE	$\mathbf{0.650 \pm 0.003}$	0.413 ± 0.002	0.376 ± 0.001	$\mathbf{0.597 \pm 0.002}$	0.384 ± 0.001	0.365 ± 0.001
LSTM	0.130 ± 0.003	0.097 ± 0.003	0.089 ± 0.002	0.593 ± 0.002	$\mathbf{0.439 \pm 0.002}$	$\mathbf{0.420 \pm 0.002}$
HRED	0.580 ± 0.003	$\mathbf{0.418 \pm 0.003}$	$\mathbf{0.384 \pm 0.002}$	$\mathbf{0.599 \pm 0.002}$	$\mathbf{0.439 \pm 0.002}$	$\mathbf{0.422 \pm 0.002}$

Table 2: Models evaluated using the vector-based evaluation metrics, with 95% confidence intervals.

Human study

- Created 100 questions each for Twitter and Ubuntu datasets (20 contexts with responses from 5 'diverse models')
- 25 volunteers from CS department at McGill
- Asked to judge response quality on a scale from 1 to 5
- Compared human ratings with ratings from automatic evaluation metrics

Models for response variety

1) Randomly selected response

2) Retrieval models:
   - Response with smallest TF-IDF cosine distance
   - Response selected by Dual Encoder (DE) model

3) Generating models:
   - Hierarchical recurrent encoder-decoder (HRED)

4) Human-written response (<u>not</u> ground truth)

Goal (inter-annotator)

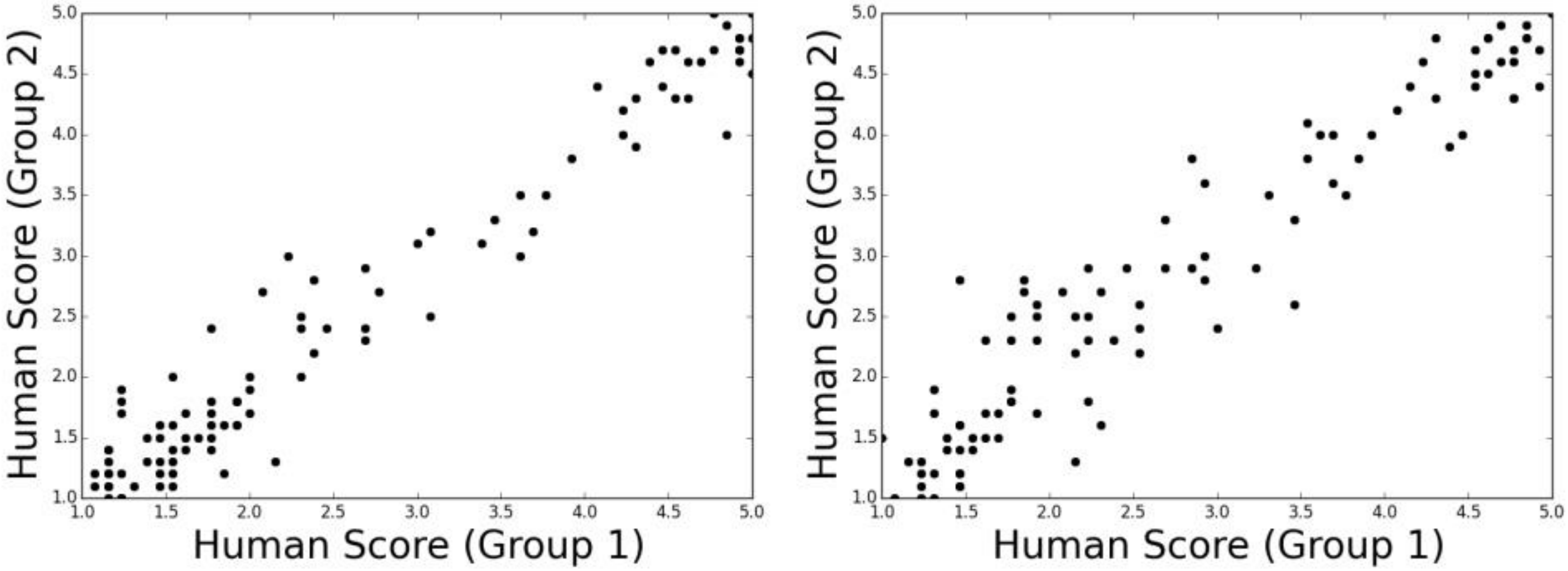
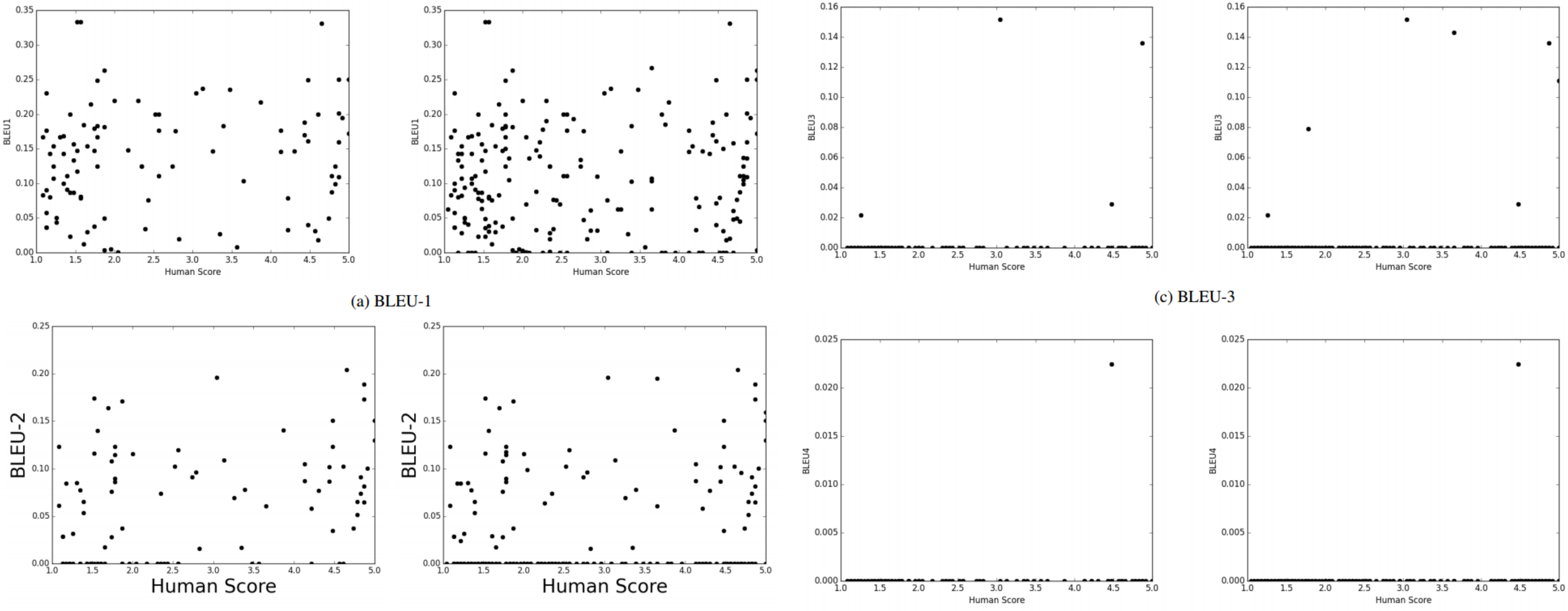


Figure 3: Scatter plots showing the correlation between two randomly chosen groups of human volunteers on the Twitter corpus (left) and Ubuntu Dialogue Corpus (right).

Reality (BLEU)



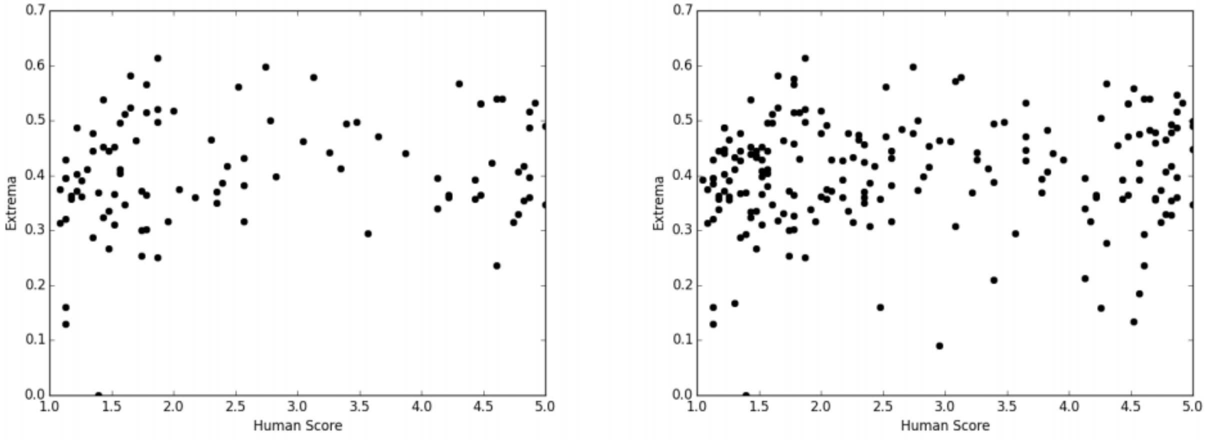
(a) BLEU-1

(b) BLEU-2

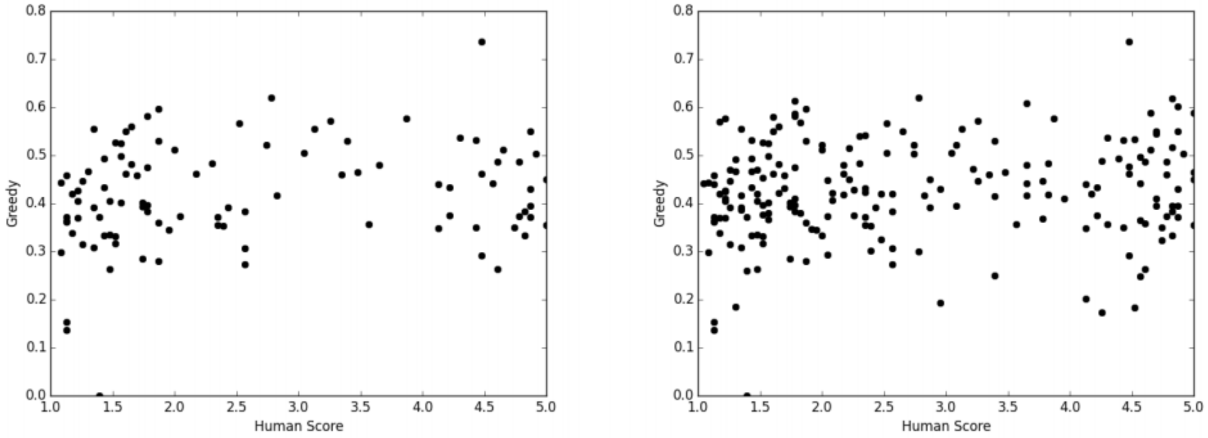
(c) BLEU-3

(d) BLEU-4

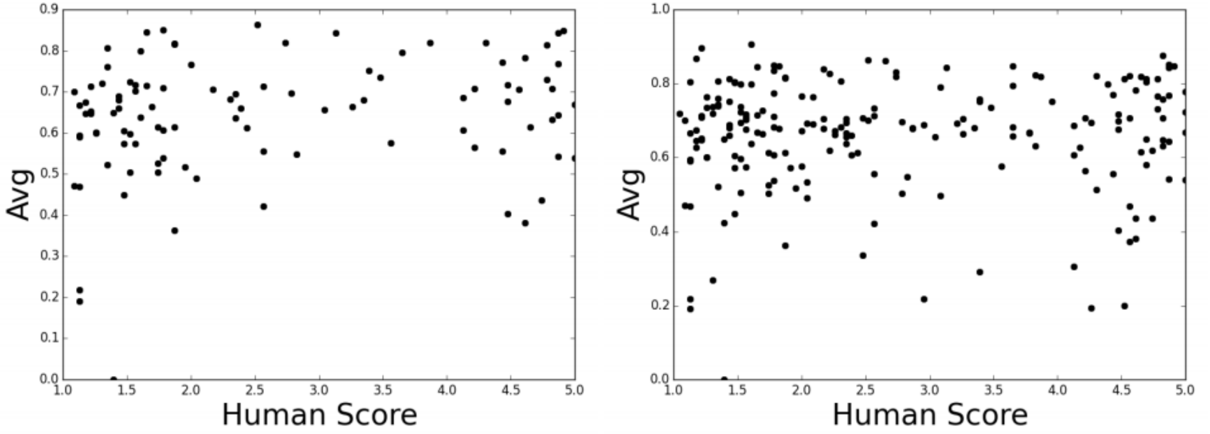
Reality (vector-based)



(a) Vector Extrema

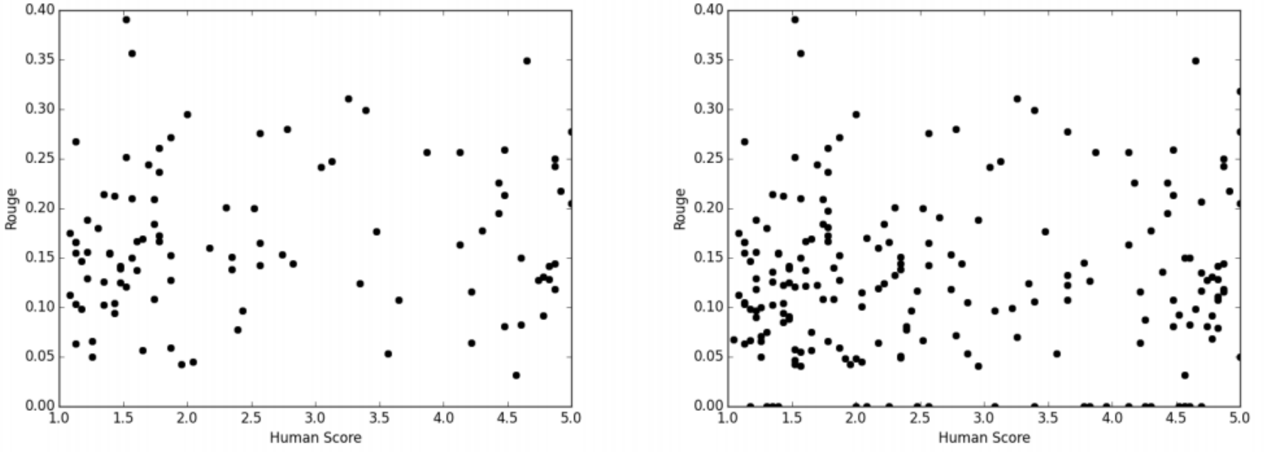


(b) Greedy Matching

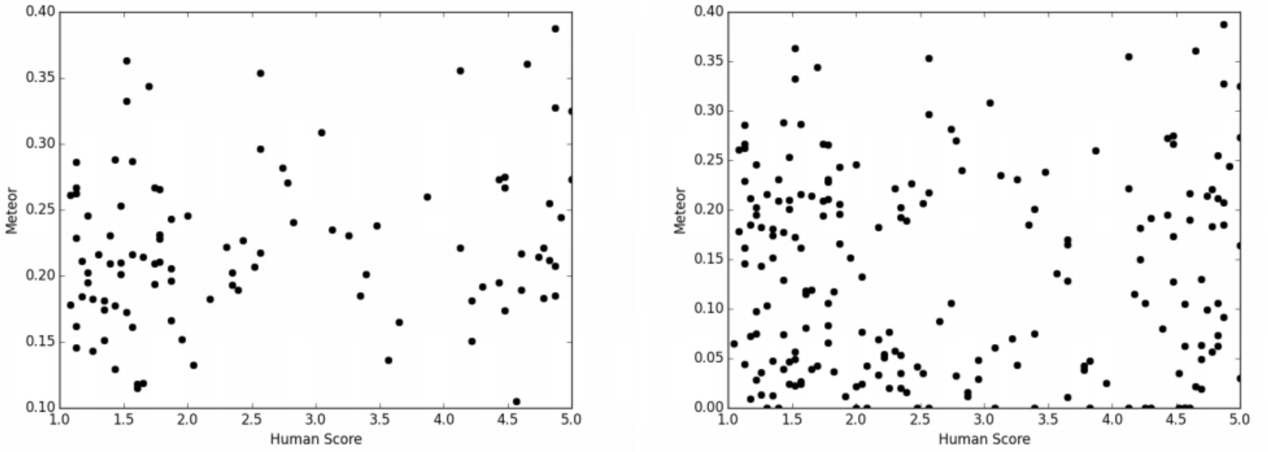


(c) Vector Averaging

Reality (ROUGE & METEOR)



(a) ROUGE



(b) METEOR

Caveats & future work

- This analysis holds when we have only one ground-truth utterance
- If you are conditioning on ‘extra information’, BLEU score might be fine
- Future work: train evaluation model on (more) human annotated data

Other curiosities

- Hard to evaluate when the proposed response has a different length than the ground-truth response

	Mean score		
	$\Delta w <= 6$ (n=47)	$\Delta w >= 6$ (n=53)	p-value
BLEU-1	0.1724	0.1009	< 0.01
BLEU-2	0.0744	0.04176	< 0.01
Average	0.6587	0.6246	0.25
METEOR	0.2386	0.2073	< 0.01
Human	2.66	2.57	0.73

Other curiosities

- Removing stop words from BLEU evaluation actually makes things worse

	Spearman	p-value	**Pearson**	p-value
BLEU-1	0.1580	0.12	0.2074	0.038
BLEU-2	0.2030	0.043	0.1300	0.20

Table 4: Correlation between BLEU metric and human judgements after removing stopwords and punctuation for the Twitter dataset.