Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses (Supplemental Material)

Anonymous ACL submission

Appendix A: Further Notes on Crowdsourcing Data Collection

Amazon Mechanical Turk Experiments We conducted two rounds of AMT experiments. We first asked AMT workers to provide a reasonable continuation of a Twitter dialogue (i.e. generate the next response given the context of a conversation). Each survey contained 20 questions, including an attention check question. Workers were instructed to generate longer responses, in order to avoid simple one-word responses. In total, we obtained approximately 2,000 human responses.

Second, we filtered these human-generated responses for potentially offensive language, and combined them with approximately 1,000 responses from each of the above models into a single set of responses. We then asked AMT workers to rate the overall quality of each response on a scale of 1 (low quality) to 5 (high quality). Each user was asked to evaluate 4 responses from 50 different contexts. We included four additional attention-check questions and a set of five contexts was given to each participant for assessment of inter-annotator agreement. We removed all users who either failed an attention check question or achieved a κ inter-annotator agreement score lower than 0.2 (Cohen, 1968). The remaining evaluators had a median κ score of 0.63, indicating moderate agreement. This is consistent with results from (Liu et al., 2016). Dataset statistics are provided in Table **??**.

In initial experiments, we also asked humans to provide scores for topicality, informativeness, and whether the context required background information to be understandable. Note that we did not ask for fluency scores, as 3/4 of the responses were produced by humans (including the retrieval models). We found that scores for informativeness and background had low inter-annotator agreement (Table 1), and scores for topicality were highly correlated with the overall score (Pearson correlation of 0.72). Results on these auxiliary questions varied depending on the wording of the question. Thus, we continued our experiments by only asking for the overall score. We provide more details concerning the data collection in the supplemental material, as it may aid others in developing effective crowdsourcing experiments.

Measurement	**κ score**
Overall	0.63
Topicality	0.57
Informativeness	0.31
Background	0.05

Table 1: Median κ inter-annotator agreement scores for various questions asked in the survey.

Preliminary AMT experiments Before conducting the primary crowdsourcing experiments to collect the dataset in this paper, we ran a series of preliminary experiments to see how AMT workers responded to different questions. Unlike the primary study, where we asked a small number of overlapping questions to determine the κ score and filtered users based on the results, we conducted a study where all responses (40 in total from 10 contexts) were overlapping. We did this for 18 users in two trials, resulting in 153 pair-wise correlation scores per trial.

In the first trial, we asked the following questions to the users, for each response:

1. How appropriate is the response overall? (overall, scale of 1-5)

2. How on-topic is the response? (topicality, scale of 1-5)

3. How specific is the response to some context? (specificity, scale of 1-5)

4. How much background information is required to understand the context? (background, scale of 1-5)

Note that we do not ask for fluency, as the 3/4 responses for each context were written by a human (including retrieval models). We also provided the AMT workers with examples that have high topicality and low specificity, and examples with high specificity and low topicality. The background question was only asked once for each context.

We observed that both the overall scores and topicality had fairly high inter-annotator agreement (as shown in Table 1), but were strongly correlated with each other (i.e. participants would often put the same scores for topicality and overall score). Conversely, specificity ($\kappa = 0.12$) and background ($\kappa = 0.05$) had very low inter-annotator agreements.

To better visualize the data, we produce scatterplots showing the distribution of scores for different responses, for each of the four questions in our survey (Figure 1). We can see that the overall and topicality scores are clustered for each question, indicating high agreement. However, these clusters are most often in the same positions for each response, which indicates that they are highly correlated with each other. Specificity and background information, on the other hand, show far fewer clusters, indicating lower inter-annotator agreement. We conjectured that this was partially because the terms 'specificity' and 'background information', along with our descriptions of them, had a high cognitive load, and were difficult to understand in the context of our survey.

To test this hypothesis, we conducted a new survey where we tried to ask the questions for specificity and background in a more intuitive manner. We also changed the formulation of the background question to be a binary 0-1 decision of whether users understood the context. We asked the following questions:

1. How appropriate is the response overall? (overall, scale of 1-5)

2. How on-topic is the response? (topicality, scale of 1-5)

3. How common is the response? (informativeness, scale of 1-5)

4. Does the context make sense? (context, scale of 0-1)

We also clarified our description for the third question, including providing more intuitive examples. Interestingly, the inter-annotator agreement on informativeness $\kappa = 0.31$ was much higher than that for specificity in the original survey. Thus, the formulation of questions in a crowdsourcing survey has a large impact on inter-annotator agreement. For the context, we found that users either agreed highly ($\kappa > 0.9$ for 45 participants), or not at all ($\kappa < 0.1$ for 113 participants).

We also experimented with asking the overall score on a separate page, before asking questions 2-4, and found that this increased the κ agreement slightly. Similarly, excluding all scores where participants indicated they did not understand the context improved inter-annotator agreement slightly.

Due to these observations, we decided to only ask users for their overall quality score for each response, as it is unclear how much additional information is provided by the other questions in the context of dialogue. We hope this information is useful for future crowdsourcing experiments in the dialogue domain.

Appendix B: Metric Description

BLEU BLEU (Papineni et al., 2002) analyzes the co-occurrences of n-grams in the ground truth and the proposed responses. It first computes an n-gram precision for the whole dataset:

$$P_n(r, \hat{r}) = \frac{\sum_k \min(h(k,r), h(k,\hat{r}_i))}{\sum_k h(k, r_i)}$$

where k indexes all possible n-grams of length n and $h(k, r)$ is the number of n-grams k in r. Note that the min in this equation is calculating the number of co-occurrences of n-gram k between the ground truth response r and the proposed response $\hat{r}$, as it computes the fewest appearances of k in either response. To avoid the drawbacks of using a precision score, namely that it favours shorter (candidate) sentences, the authors introduce a brevity penalty. BLEU-N, where N is the maximum length of n-grams considered, is defined as:

$$\text{BLEU-N} := b(r, \hat{r}) \exp(\sum_{n=1}^{N} \beta_n \log P_n(r, \hat{r}))$$

β_n is a weighting that is usually uniform, and $b(\cdot)$ is the brevity penalty. The most commonly used version of BLEU assigns $N = 4$. Modern versions of BLEU also use sentence-level smoothing, as

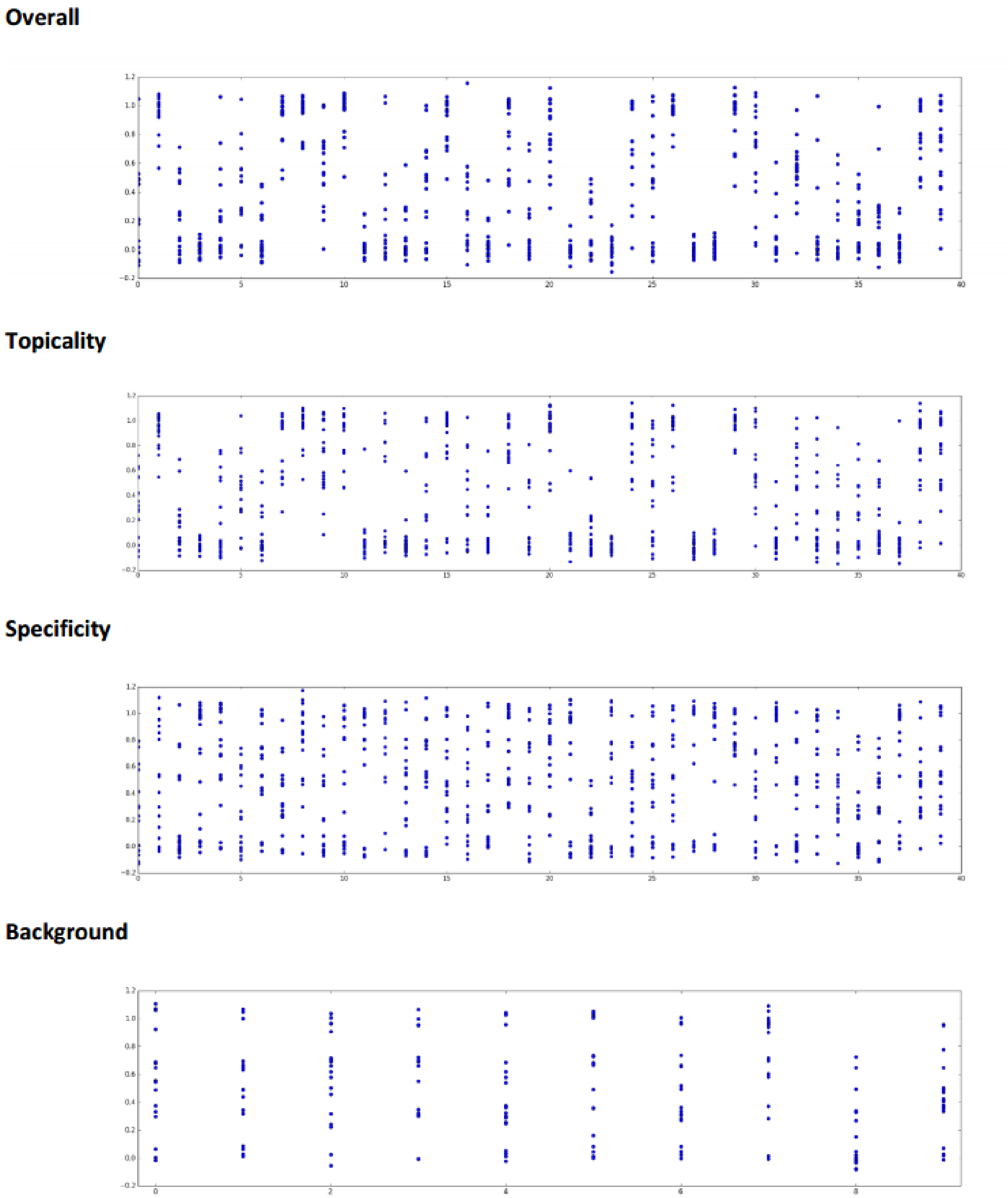


Figure 1: Scatter plots showing the distribution of scores (vertical axis) for different responses (horizontal axis), for each of the four questions in our survey. It can be seen that the overall and topicality scores are clustered for each question, indicating high agreement, while this is not the case for specificity or background information. Note that all scores are normalized based on a per-user basis, based on the average score given by each user.

the geometric mean often results in scores of 0 if there is no 4-gram overlap (Chen and Cherry, 2014). Note that BLEU is usually calculated at the corpus-level, and was originally designed for use with multiple reference sentences.

METEOR The METEOR metric (Banerjee and Lavie, 2005) was introduced to address several weaknesses in BLEU. It creates an explicit alignment between the candidate and target responses. The alignment is based on exact token matching, followed by WordNet synonyms, stemmed tokens, and then paraphrases. Given a set of alignments, the METEOR score is the harmonic mean of precision and recall between the proposed and ground truth sentence.

Given a set of alignments m, the METEOR score is the harmonic mean of precision P_m and recall R_m between the candidate and target sentence.

$$Pen = \gamma(\frac{ch}{m})^{\theta} \tag{1}$$

$$F_{mean} = \frac{P_m R_m}{\alpha P_m + (1-\alpha) R_m} \tag{2}$$

$$P_m = \frac{|m|}{\sum_k h_k(c_i)} \tag{3}$$

$$R_m = \frac{|m|}{\sum_k h_k(s_{ij})} \tag{4}$$

$$METEOR = (1 - Pen) F_{mean} \tag{5}$$

The penalty term Pen is based on the 'chunkiness' of the resolved matches. We use the default values for the hyperparameters α, γ, and θ.

ROUGE ROUGE (Lin, 2004) is a set of evaluation metrics used for automatic summarization. We consider ROUGE-L, which is a F-measure based on the Longest Common Subsequence (LCS) between a candidate and target sentence. The LCS is a set of words which occur in two sentences in the same order; however, unlike n-grams the words do not have to be contiguous, i.e. there can be other words in between the words of the LCS. ROUGE-L is computed using an F-measure between the reference response and the proposed response.

$$R = \max_j \frac{l(c_i, s_{ij})}{|s_{ij}|} \tag{6}$$

$$P = \max_j fracl(c_i, s_{ij})|c_{ij}| \tag{7}$$

$$ROUGE_L(c_i, S_i) = \frac{(1+\beta^2)RP}{R + \beta^2 P} \tag{8}$$

where $l(c_i, s_{ij})$ is the length of the LCS between the sentences. β is usually set to favour recall ($\beta = 1.2$).

Appendix C: Latent Variable Hierarchical Recurrent Encoder-Decoder (VHRED)

The VHRED model is an extension of the original hierarchical recurrent encoder-decoder (HRED) model (Serban et al., 2016) with an additional component: a high-dimensional stochastic latent variable at every dialogue turn. The dialogue context is encoded into a vector representation using the *utterance-level* and *context-level* RNNs from our encoder. Conditioned on the summary vector at each dialogue turn, VHRED samples a multivariate Gaussian variable that is provided, along with the context summary vector, as input to the *decoder* RNN, which in turn generates the response word-by-word. We use representations from the VHRED model as it produces more diverse and coherent responses compared to its HRED counterpart.

The VHRED model is trained to maximize a lower-bound on the log-likelihood of generating the next response:

$$\begin{aligned}\mathcal{L} &= \log P_{\hat{\theta}}(\mathbf{w}_1, \ldots, \mathbf{w}_N) \\ &\geq \sum_{n=1}^{N} -\mathrm{KL}\left[Q_\psi(\mathbf{z}_n \mid \mathbf{w}_1, \ldots, \mathbf{w}_n) || P_{\hat{\theta}}(\mathbf{z}_n \mid \mathbf{w}_{<n})\right] \\ &+ \mathbb{E}_{Q_\psi(\mathbf{z}_n|\mathbf{w}_1,\ldots,\mathbf{w}_n)}\left[\log P_{\hat{\theta}}(\mathbf{w}_n \mid \mathbf{z}_n, \mathbf{w}_{<n})\right],\end{aligned} \tag{9}$$

where $\mathrm{KL}[Q||P]$ is the Kullback-Leibler (KL) divergence between distributions Q and P. The distribution $Q_\psi(\mathbf{z}_n \mid \mathbf{w}_1, \ldots, \mathbf{w}_N) = \mathcal{N}(\boldsymbol{\mu}_{\text{posterior}}(\mathbf{w}_1, \ldots, \mathbf{w}_n), \Sigma_{\text{posterior}}(\mathbf{w}_1, \ldots, \mathbf{w}_n))$ is the approximate posterior distribution (or *recognition model*) which approximates the intractable true posterior distribution $P_\psi(\mathbf{z}_n \mid \mathbf{w}_1, \ldots, \mathbf{w}_N)$. The posterior mean $\boldsymbol{\mu}_{\text{posterior}}$ and covariance $\Sigma_{\text{posterior}}$ (as well as that of the prior) are computed using a feed-forward neural network, which takes as input the concatenation of the vector representations of the past utterances and that of the current utterance.

The multivariate Gaussian latent variable in the VHRED model allows modelling ambiguity and uncertainty in the dialogue through the latent variable distribution parameters (mean and variance). This provides a useful inductive bias, which helps VHRED encode the dialogue context into a real-valued embedding space even when the dialogue context is ambiguous or uncertain, and it helps VHRED generate more diverse responses.

Pre-training motivation Maximizing the likelihood of generating the next utterance in a dialogue is not only a convenient way of training the encoder parameters; it is also an objective that is consistent with learning useful representations of the dialogue utterances. Two context vectors produced by the VHRED encoder are similar if the contexts induce a similar distribution over subsequent responses; this is consistent with the formulation of the evaluation model, which assigns high scores to responses that have similar vector representations to the context. VHRED is also closely related to the skip-thought-vector model (Kiros et al., 2015), which has been shown to learn useful representations of sentences for many tasks, including semantic relatedness and paraphrase detection. The skip-thought-vector model takes as input a single sentence and predicts the previous sentence and next sentence. On the other hand, VHRED takes as input several consecutive sentences and predicts the next sentence. This makes it particularly suitable for learning long-term context representations.

Appendix D: Experiments & results

Hyperparameters

When evaluating our model, we conduct early stopping on an external validation set to obtain the best parameter setting. We similarly choose our hyperparameters (PCA dimension n, L2 regularization penalty γ, learning rate a, and batch size b) based on validation set results. Our best ADEM model used $\gamma = 0.075$, $a = 0.01$, and $b = 32$. For ADEM with tweet2vec embeddings, we did a similar hyperparameter searched, and used $n = 150$, $\gamma = 0.01$, $a = 0.01$, and $b = 16$.

Additional Results

New results on (Liu et al., 2016) data In order to ensure that the correlations between word-overlap metrics and human judgements were comparable across datasets, we standardized the processing of the evaluation dataset from (Liu et al., 2016). In particular, the original data from (Liu et al., 2016) has a token (either '<first_speaker>', '<second_speaker>', or '<third_speaker>') at the beginning of each utterance. This is an artifact left-over by the processing used as input to the hierarchical recurrent encoder-decoder (HRED) model

Metric	Spearman	Pearson
BLEU-1	-0.026 (0.80)	0.016 (0.87)
BLEU-2	0.065 (0.52)	0.080 (0.43)
BLEU-3	0.139 (0.17)	0.088 (0.39)
BLEU-4	0.139 (0.17)	0.092 (0.36)
ROUGE	-0.083 (0.41)	-0.010 (0.92)

Table 2: Correlations between word-overlap metrics and human judgements on the dataset from (Liu et al., 2016), after removing the speaker tokens at the beginning of each utterance. The correlations are even worse than estimated in the original paper, and none are significant.

Metric	Wall time
ADEM (CPU)	2861s
ADEM (GPU)	168s

Table 3: Evaluation time on the test set.

(Serban et al., 2016). Removing these tokens makes sense for establishing the ability of word-overlap models, as they are unrelated to the content of the tweets.

We perform this processing, and report the updated results for word-overlap metrics in Table 2. Surprisingly, almost all significant correlation disappears, particularly for all forms of the BLEU score. Thus, we can conclude that the word-overlap metrics were heavily relying on these tokens to form bigram matches between the model responses and reference responses.

Evaluation speed An important property of evaluation models is speed. We show the evaluation time on the test set for ADEM on both CPU and a Titan X GPU (using Theano, without cudNN) in Table 3. When run on GPU, ADEM is able to evaluate responses in a reasonable amount of time (approximately 2.5 minutes). This includes the time for encoding the contexts, model responses, and reference responses into vectors with the hierarchical RNN, in addition to computing the PCA projection, but does not include pre-training with VHRED. For comparison, if run on a test set of 10,000 responses, ADEM would take approximately 45 minutes. This is significantly less time consuming than setting up human experiments at any scale. Note that we have not yet made any effort to optimize the speed of the ADEM model.

Learning curves To show that our learning procedure for ADEM really is necessary, and that the embeddings produced by VHRED are not sufficient to evaluate dialogue systems, we plot the Spearman and Pearson correlations on the test set as a function of the number of epochs in Figure 2. It is clear that, at the beginning of training, when the matrices M and N have been initialized to the identity, the model is incapable of accurately predicting human scores, and its correlation is approximately 0.

Failure analysis We now conduct a failure analysis of the ADEM model. In particular, we look at two different cases: responses where both humans and (normalized) ROUGE or BLEU-2 score highly (a score of 4 out of 5 or greater) while ADEM scores poorly (2 out of 5 or lower), and the converse, where ADEM scores the response highly while humans and either ROUGE or BLEU-2 score it poorly. We randomly sample (i.e. without cherry picking) three examples of each case, which are shown in Tables 4-5.

From Table 4, the cases where ADEM misses a good response, we can see that there are a variety of reasons for this cause of failure. In the first example, ADEM is not able to match the fact that the model response talks about sleep to the reference response or context. This is possibly because the utterance contains a significant amount of irrelevant information: indeed, the first two sentences are not related to either the context or reference response. In the second example, the model response does not seem particularly relevant to the context — despite this, the human scoring this example gave it $4/5$. This illustrates one drawback of human evaluations; they are quite subjective, and often have some noise. This makes it difficult to learn an effective ADEM model. Finally, ADEM is unable to score the third response highly, even though it is very closely related to the reference response.

We can observe from the first two examples in Table 5, where the ADEM model erroneously ranks the model responses highly, that ADEM is occasionally fooled into giving high scores for responses that are completely unrelated to the context. This may be because both of the utterances are short, and short utterances are ranked higher by humans in general since they are often more generic (as detailed in Section **??**). In the third example, the response actually seems to be somewhat reasonable given the context; this may be an instance where the human evaluator provided a score that was too low.

Data efficiency How much data is required to train ADEM? We conduct an experiment where we train ADEM on different amounts of training data, from 5% to 100%. The results are shown in Table 6. We can observe that ADEM is very data-efficient, and is capable of reaching a Spearman correlation of 0.4 using only half of the available training data (1000 labelled examples). ADEM correlates significantly with humans even when only trained on 5% of the original training data (100 labelled examples).

Improvement over word-overlap metrics Next, we analyze more precisely how ADEM outperforms traditional word-overlap metrics such as BLEU-2 and ROUGE. We first normalize the metric scores to have the same mean and variance as human scores, clipping the resulting scores to the range $[1, 5]$ (we assign raw scores of 0 a normalized score of 1). *We indicate normalization with vertical bars around the metric.* We then select all of the good responses that were given low scores by word-overlap metrics (i.e. responses which humans scored as 4 or higher, and which |BLEU-2| and |ROUGE| scored as 2 or lower). The results are summarized in Table 7: of the 237 responses that humans scored 4 or higher, most of them (147/237) were ranked very poorly by both BLEU-2 and ROUGE. This quantitatively demonstrates what we argued qualitatively in Figure **??**; a major failure of word-overlap metrics is the inability to consider reasonable responses that have no word-overlap with the reference response. We can also see that, in almost half (60/147) of the cases where both BLEU-2 and ROUGE fail, |ADEM| is able to correctly assign a score greater than 4. For comparison, there are only 42 responses where humans give a score of 4 and |ADEM| gives a score less than 2, and only 14 of these are assigned a score greater than 4 by either |BLEU-2| or |ROUGE|.

To provide further insight, we give specific examples of responses that are scored highly (> 4) by both humans and |ADEM|, and poorly (< 2) by both |BLEU-2| and |ROUGE| in Table 9. We draw 3 responses randomly (i.e. no cherry-picking) from the 60 test set responses that meet this criteria. We can observe that ADEM is able to recognize short responses that are appropriate to the context, without word-overlap with the reference response. This is even the case when the model and reference responses have very little semantic similarity, as in the first and third examples in Table 9.

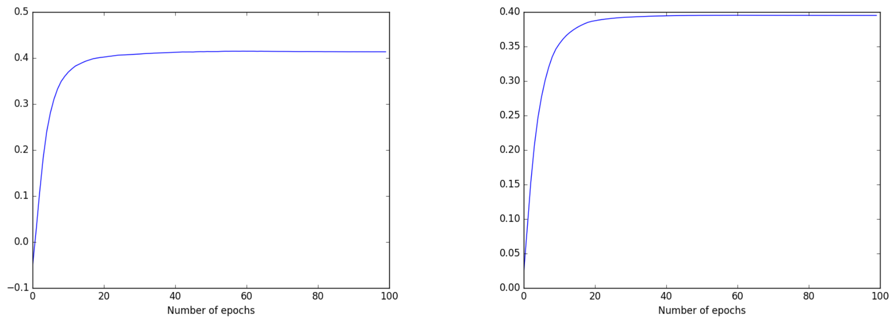
Finally, we show the behaviour of ADEM when

Context	Reference response	Model response	Human score	|BLEU-2| score	|ROUGE| score	|ADEM| score
what theme do you guys want next on tumblr? we've had mariskamommymoments what do you want to see next? → im sorry. hope you feel better soon! → it will wear off. just hate feeling like this → im sure it will! just relax and take your time → i'm okay. just overly tired	maybe you need a bit more sleep	i'm really impressed. first guy to said that p what's time? sleep late its not good. i'm worried	4	2.53	5.0	1.0
1 some pm syria - the editor of syrian government daily tishrin said she had been sacked over remarks to al <url> → replaced by business editor! sounds like the states lol	msm is very similar the world over ! tied to the government .	i'm not sure if i'm going to be able to see the <unk>. i'm not sure if i should be able to see it.	4	2.53	4.75	1.22
wonder how long it will take wayne rooney to get himself into trouble on twitter. odds? → probably. a long time. because i used the address in the paper to find the page and it doesn't exist! → here you go	thanks . i wonder why it didn't work !	thanks let me see if this one works :p	5	2.53	4.24	1.53

Table 4: Examples where a human and either BLEU-2 or ROUGE (after normalization) score the model response highly ($> 4/5$), while the ADEM model scored it poorly ($< 2/5$). These examples are drawn randomly (i.e. no cherry-picking). The bars around |metric| indicate that the metric scores have been normalized.

Context	Reference response	Model response	Human score	|BLEU-2| score	|ROUGE| score	|ADEM| score
rage slightly dissipated. wouldn't have bothered restoring my phone but i need it to moan at tomorrow. → speaking of moaning. i'm actually going to email that chap that letter right now. → good plan	i 'm going to do a little wee blog about it too . all nice and measured , of course .	some. some unfortunately	2	2.53	1.57	4.38
high school flings college relationships → it seems like the other way around from wat i've seen	word . i 've seen a little of both . more of the college though	king james	1	2.53	1.57	5.0
is it getting light outside? i swear it looks blue. → time to go to sleepppp..	for you , i 'm staying up	i'm going to the beach.	1	2.53	1.57	5.0

Table 5: Examples where a human and either BLEU-2 or ROUGE (after normalization) score the model response low ($< 2/5$), while the ADEM model scored it highly ($> 4/5$). These examples are drawn randomly (i.e. no cherry-picking). The bars around |metric| indicate that the metric scores have been normalized.



(a) Spearman correlation

(b) Pearson correlation

Figure 2: Plots showing the Spearman and Pearson correlations on the test set as ADEM trains. At the beginning of training, the model does not correlate with human judgements.

Training data %	**Spearman**	p-value	**Pearson**	p-value
100 % of data	0.414	< 0.001	0.395	< 0.001
75 % of data	0.408	< 0.001	0.393	< 0.001
50 % of data	0.400	< 0.001	0.391	< 0.001
25 % of data	0.330	< 0.001	0.331	< 0.001
10 % of data	0.245	< 0.001	0.265	< 0.001
5 % of data	0.098	0.015	0.161	< 0.001

Table 6: ADEM correlations when trained on different amounts of data.

Metric scores	**# Examples**
Human ≥ 4	237 out of 616
and (|BLEU-2| <2, |ROUGE| <2)	146 out of 237
and |ADEM| > 4	60 out of 146
and |ADEM| < 2	42 out of 237
and (|BLEU-2| >4, **or** |ROUGE| >4)	14 out of 42

Table 7: In 60/146 cases, ADEM scores good responses (human score > 4) highly when word-overlap metrics fail. The bars around |metric| indicate that the metric scores have been normalized.

	Mean score		
	$\Delta w \leq 6$ (n=312)	$\Delta w > 6$ (n=304)	**p-value**
ROUGE	0.042	0.031	< 0.01
BLEU-2	0.0022	0.0007	0.23
ADEM	2.072	2.015	0.23
Human	2.671	2.698	0.83

Table 8: Effect of differences in response length on the score, Δw = absolute difference in #words between the reference response and proposed response. BLEU-1, BLEU-2, and METEOR have previously been shown to exhibit bias towards similar-length responses (Liu et al., 2016).

there is a discrepancy between the lengths of the reference and model responses. In (Liu et al., 2016), the authors show that word-overlap metrics such as BLEU-1, BLEU-2, and METEOR exhibit a bias in this scenario: they tend to assign higher scores to responses that are closer in length to the reference response.[1] However, humans do not exhibit this bias; in other words, the quality of a response as judged by a human is roughly independent of its length. In Table 8, we show that ADEM also does not exhibit this bias towards similar-length responses.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. volume 29, pages 65–72.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. *ACL 2014* page 362.

[1]Note that, for our dataset, BLEU-2 almost exclusively assigns scores near 0 for both $\Delta w \leq 6$ and $\Delta w > 6$, resulting in a p-value >0.05.

Context	Reference response	Model response	Human score	|BLEU-2| score	|ROUGE| score	|ADEM| score
i'd recommend <url> - or build buy an htpc and put <url> on it. → you're the some nd person this week that's recommended roku to me.	an htpc with xmbc is what i run . but i 've decked out my setup . i 've got <number> tb of data on my home server	because it's brilliant	5	1.0	1.0	4.726
imma be an auntie this weekend. i guess i have to go albany. herewego → u supposed to been here → i come off nd on. → never tell me smh	lol you sometiming	haha, anyway, how're you?	5	1.0	1.0	4.201
my son thinks she is plain. and the girl that plays her sister. seekhelp4him? → send him this. he'll thank you. <url>	you are too kind for words .	i will do	5	1.0	1.0	5.0

Table 9: Examples where both human and ADEM score the model response highly, while BLEU-2 and ROUGE do not. These examples are drawn randomly (i.e. no cherry-picking) from the examples where ADEM outperforms BLEU-2 and ROUGE (as defined in the text). ADEM is able to correctly assign high scores to short responses that have no word-overlap with the reference response. The bars around |metric| indicate that the metric scores have been normalized.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70(4):213.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*. pages 3276–3284.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*. Barcelona, Spain, volume 8.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023* .

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 311–318.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*. pages 3776–3784.