Re-Evaluating ADEM: A Deeper Look at Scoring Dialogue Responses

Ananya B. Sai,[1,2,4] **Mithun Das Gupta,**[3] **Mitesh M. Khapra,**[1,2] **Mukundhan Srinivasan**[4]
[1]Department of Computer Science and Engineering, Indian Institute of Technology, Madras
[2]Robert Bosch Center for Data Sciences and AI (RBC-DSAI), Indian Institute of Technology, Madras
[3]Microsoft, India
[4]NVIDIA, India
{ananyasb,miteshk}@cse.iitm.ac.in, migupta@microsoft.com, msrinivasan@nvidia.com

Abstract

Automatically evaluating the quality of dialogue responses for unstructured domains is a challenging problem. ADEM (Lowe et al. 2017) formulated the automatic evaluation of dialogue systems as a learning problem and showed that such a model was able to predict responses which correlate significantly with human judgements, both at utterance and system level. Their system was shown to have beaten word-overlap metrics such as BLEU with large margins. We start with the question of whether an adversary can game the ADEM model. We design a battery of targeted attacks at the neural network based ADEM evaluation system and show that automatic evaluation of dialogue systems still has a long way to go. ADEM can get confused with a variation as simple as reversing the word order in the text! We report experiments on several such adversarial scenarios that draw out counterintuitive scores on the dialogue responses. We take a systematic look at the scoring function proposed by ADEM and connect it to linear system theory to predict the shortcomings evident in the system. We also devise an attack that can fool such a system to rate a response generation system as favorable. Finally, we allude to future research directions of using the adversarial attacks to design a truly automated dialogue evaluation system.

Introduction

AI agents capable of having human-like conversations find various applications such as providing an automated helpdesk for customer service and technical support, serving as language learning tools, personal assistants and as a source of entertainment/recreation. The research community has accessibility to a number of datasets for the task of dialogue generation (Ritter, Cherry, and Dolan 2010; Lowe et al. 2015; Saha, Khapra, and Sankaranarayanan 2018). This has led to the emergence of goal-driven as well as non-goal-driven conversation models (Ritter, Cherry, and Dolan 2011; Vinyals and Le 2015; Wen et al. 2015; Sordoni et al. 2015; Yu et al. 2017; Li et al. 2017). However it is hard to measure the scientific progress towards a conversational agent due to the lack of good evaluation metrics. Human evaluations and comparisons are reported in most of the works on samples of the data. However, it is infeasible to have a human in the loop to give feedback and scores for training and evaluating dialogue systems. This has led to adoption of the existing automatic evaluation metrics for the task of scoring the generated dialogues. Popular word overlap based metrics such as BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE (Lin 2004) and the various word embedding based metrics have been used to score dialogue generation systems. However, they can't handle the diversity of the range of valid responses and are shown to correlate poorly with human judgement (Liu et al. 2016).

Context:	Do you want to watch a movie today?
Valid responses:	Sure, Avengers infinity war is out. I didn't finish my assignment. Yeah, but will it rain tonight?

Table 1: Diversity of valid responses in a Dialogue

Consider the conversation in Table 1. All of the responses sound reasonable and logical given the context. However, one can notice that there is no word overlap across any pair of the responses. Should any one of the responses be considered the ground truth, a model that produces any other response will receive very low scores. The popular BLEU metric would assign the responses a score of 0. Even the word embeddings based metrics face a similar issue. The task of dialogue generation is tough to evaluate due to the huge number of valid responses possible.

Simultaneously research efforts are underway in the direction of using generative adversarial network (GAN) based dialogue systems where the generator and discriminator are trained in parallel. The discriminator's role is to determine whether the given dialogue is machine-generated or human-generated. However, it is tricky to use the discriminator trained this way as one cannot be sure if it is infact a good discriminator or seems good due to a poorer generator. Such training might also lead to a very model specific evaluator rather than a generic one.

In this work we look at one such recently proposed metric, viz. Automatic Dialogue Evaluation Model (ADEM) by Lowe et al. (2017) which trains a neural net based model to score the overall quality of dialogue responses on a scale of 1-5. ADEM is trained on Twitter corpus with the help of AMT (Amazon Mechanical Turk) for obtaining human scores for the dialogues. ADEM uses a hierarchical RNN en-

coder architecture to obtain the context embedding as well as the embeddings of the model response (*i.e.*, the response to be evaluated) and the reference response. The score of the response is a function of these embeddings. The primary flaw with this is the multiplicative setting of embedding interactions which has been shown to result in embedding vectors of high conicity (lesser spread) by Chandrahas et al. (2018). We provide an analysis of this phenomenon and verify the claim experimentally as well.

We do a deeper analysis and show that the scores assigned by ADEM to various dialog responses are banded in a small margin around the mean of about 2.75 with a spread of around 0.34. Further analysis shows that due to the inherent flaw in its design, ADEM is actually susceptible to adversarial attacks wherein it fails to provide appropriate scores for multiple responses. This is counterintuitive to the claims made by ADEM, which tries to find diversified scores for varied responses correlated to human scores. Lastly, we look at the cost function of ADEM and prove that the spread cannot be widened by the multiplicative model leading to a high conicity. ADEM was proposed with the motivation of addressing the limitations of metrics such as BLEU by providing high scores to diverse responses correlating human scores, but in the process it scores most of the responses extremely closely and shows very low discriminative power to handle a wide array of real life responses.

Related Work

Since our work focuses on a critique of automatic evaluation metrics we first do a quick review of various popular metrics used for automatic evaluation and then review works which are similar in idea to ours and themselves do a critique of these evaluation metrics. The research on dialogue generation models is guided by the dialogue evaluation metrics which provide the means for comparison. BLEU and METEOR scores, originally used for machine translation, are adopted for this task by various works (Yu et al. 2017; Ritter, Cherry, and Dolan 2011; Sordoni et al. 2015; Wen et al. 2015; Galley et al. 2015; Li et al. 2016a). BLEU analyses the co-occurrences of n-grams whereas METEOR creates an explicit alignment using exact matching, followed by WordNet synonyms, stemmed tokens, and paraphrases, in that order. Similarly the ROUGE metric variants, originally used for automatic summarization, work on overlapping units such as n-grams, word sub-sequences and word pairs. The ROUGE metrics being recall-oriented, require a sufficient number of references to produce reliable scores.

While the above metrics directly use words for comparison, another alternative is to use word embeddings. Word embeddings are calculated by methods such as Word2Vec (Mikolov et al. 2013), GloVe (Pennington, Socher, and Manning 2014), which represent words as vectors derived from the contexts they appear in. Using these word vectors, the *Greedy Matching metric* considers each token in actual response and greedily matches it with each token in predicted response based on cosine similarity of word embedding (and vice-versa). The total score is then averaged across all words, making this greedy approach favor responses with keywords that are semantically similar to those in the ground truth response. Certain metrics compute sentence embeddings using the word embeddings. The *Embedding Average metric* calculates sentence-level embeddings simply by averaging the word embeddings for each token/word in the sentence. Another way to calculate sentence-level embeddings is by using vector extrema, where for each dimension in the word vector, the most extreme value amongst all word vectors in the sentence is used in the sentence-level embedding. The basic idea is that by taking the maximum along each dimension, we can ignore the common words (which will be pulled towards the origin) and prioritize informative words which will lie further away in the vector space. Both these methods compare the ground truth response and the retrieved response by computing the cosine similarity between their respective sentence level embeddings.

These metrics face a lot of criticism for use in NLG tasks (Nema and Khapra 2018; Callison-Burch, Osborne, and Koehn 2006; Callison-Burch 2009). Another major shortcoming of these evaluation metrics was studied by Liu et al. (2016), who show that the scores generated by these metrics correlate poorly with human judgement. They perform this analysis using Twitter dataset consisting of casual chit-chat and Ubuntu corpus containing technical conversations and conclude that metrics which do not specifically correlate with human judgements on a new task should not be used to evaluate that task. More recently, Gao, Galley, and Li (2018) add to this debate suggesting the comparison of correlation to be made at the corpus level rather than sentence level.

	Positive / Negative ex	Ideal accuracy
1	human / human response	50%
2	machine / machine response	50%
3	human / random response	100%
4	human / skipped response	100%

Table 2: Simple scenarios for computing ERE (Evaluator Reliability Error) (Li et al. 2017)

Li et al. (2017) discuss various evaluator models and architectures for dialogue evaluation including SVM, concatenation based neural network, and a hierarchical neural network. The task of the dialogue evaluator in these cases is to classify a response as human or machine-generated. Additionally, to test the reliability of an evaluator, they introduce four scenarios as shown in Table. 2, where one can know in advance how a perfect evaluator would behave. They define a score called Evaluator Reliability Error (ERE), as the average deviation of the evaluator model from the gold standard accuracies in each of these scenarios, with equal weightage to each of the four cases. However, they evaluate their dialogue generator on the hierarchical model alone, on which they report ERE as 0.193, with their generator being able to fool the evaluator 9.8% of the times.

In this work we evaluate ADEM which scores dialogue responses and design a battery of experiments, which not only bring out the shortcomings in ADEM, but also pave

the way for re-designing such a system. We perform both theoretical and empirical evaluation of ADEM and conclude that the response evaluation systems need to address the linear system constraints as well as the semantic constraints to be uniformly useful across different dialogue domains. To the best of our knowledge, this kind of analysis of dialogue scoring systems has not been done before.

ADEM: Background

The automatic scoring method, ADEM, proposed by Lowe et al. (2017) computes the score for a dialogue response by using a dot-product between the vector representations of the dialogue context c, reference response r, and model response $\hat{r}$

$$score(c, r, \hat{r}) = (c^T M \hat{r} + r^T N \hat{r} - \alpha)/\beta \quad (1)$$

where $M, N \in \mathbb{R}^n$ are learned matrices initialized to the identity and α, β are scalar constants used to initialize the model's predictions in the range [1,5].

The model is trained to minimize the squared error between the model predictions and the human score with L2-regularization.

$$\mathcal{L} = \sum_{i=1:K} [score(c_i, r_i, \hat{r}_i) - human_i]^2 + \gamma||\theta||_2 \quad (2)$$

Using the above formulation, the authors report a high correlation between the ADEM scores and human scores at utterance level and system level. However a closer analysis of the model uncovers multiple adversarial scenarios where the evaluator is easily fooled or does the opposite of what is expected. We show that the neural network based ADEM model does not perform well with respect to the syntax and semantics of the response to be scored. The model is not able to disambiguate whether the words in a response are jumbled up or if random words of the response have been repeated. We explore the possibility of adversarial attacks and fooling mechanisms on dialogue evaluator models and propose a systematic approach to target dialogue evaluation systems. Using our method, we find irrelevant responses for the given context which ADEM scores high, concluding that ADEM is susceptible to adversarial attacks leading to significant loss in its scoring ability.

Our experiments show a consistent tendency in ADEM to score all dialogues very close to a mean value of 2.75 on a 5-point scale with 0.34 standard deviation. Further, we compute the Conicity (defined as the mean of the cosine similarities of each of the vectors with the mean vector (Chandrahas, Sharma, and Talukdar 2018)) of the response embeddings computed by ADEM. On a dataset of 200k such responses, the conicity value of the response embeddings generated by ADEM is ≈ 0.6 indicating a very low vector spread in the embedding space. Exploring further for explanations for this, we look into ADEM's score formulation and highlight our observations. Through our analysis, we intend to provide guidelines for evolution of future dialogue evaluation metrics.

Response to be evaluated	ADEM mean	ideal score
ground-truth response	2.75	5
context repeated as response	3.03	1
machine generated response*	2.64	1
swapping reference response and machine response	2.6	5

Table 3: ADEM scores on simple test cases (*Machine generated responses are obtained by training a GAN based neural dialogue generation model (Li et al. 2017))

Deeper Look into ADEM's Scoring Function and the Dialogue Embeddings

Borrowing the idea of the ERE metric proposed by Li et al. (2017), we evaluate ADEM on certain similar straightforward cases with easy-to-guess desired outcomes. While the ERE metric deals with a binary classification of dialogues as human versus machine-generated, we adopt the idea for a score based evaluators like ADEM. This analysis is presented in Table 3 using the Microsoft Research Social Media Conversation Corpus by Sordoni et al. (2015). We first check ADEM's evaluations of the reference responses. One would expect the scores on good reference response to be high. Note that the reference responses for dialogues in this dataset have high human ratings. Similarly, when the context itself is repeated back as the response, it forms an unnatural reply in most cases and one can expect scores to be very low. Next we check ADEM's scores in the case of the machine generated responses. We train a GAN based neural dialogue generation model based on Li et al. (2017) for a very limited number of 5 epochs before using the generated responses for evaluation. We deem this training to be insufficient and hence expect an ideal evaluator's scores to be low. Finally we think of the case where the reference response and the responses-to-be-evaluated (in this case, the machine generated responses) are swapped. The idea is to find out how much the reference response influences the evaluation of a response. In the numerous dialogue datasets which often involve some automation in their creation, it would not be uncommon to find certain mediocre reference responses. A human evaluator merely uses the reference response as an unabiding guideline and can recognize a better-suited response without being misled/distracted by a bad reference response. We hence suggest an ideal score of 5 for this scenario. In all of these cases, we find ADEM's score to be quite similar. We find that ADEM has a mean score between $2.6 - 3$ in all of these cases with a very low standard deviation ranging between $0.32 - 0.36$. We also find $60 - 71\%$ of the scores within 1 standard deviation of the mean.

We perform analysis to explain the possible reasons for the clustering of the scores around the mean as produced by ADEM. To proceed further with our analysis we borrow some of the concepts defined by Chandrahas et al. (2018). They define two broad classes of models which encode the behavior of embedding vectors, namely:

- multiplicative systems
- additive systems

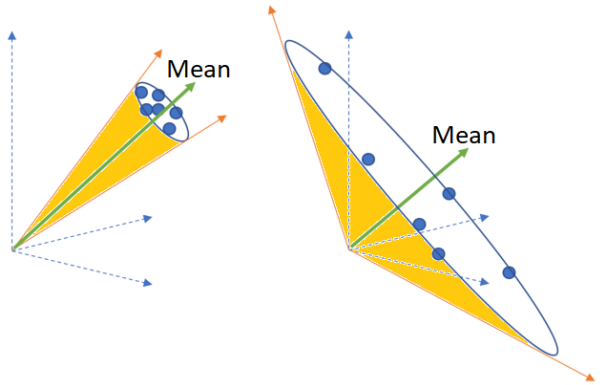

Figure 1: Conicity for a set of vectors. Left: high conicity with a small vector spread obtained from multiplicative systems. Right: low conicity with a large vector spread obtained from additive systems.

For a triplet of embedding vectors, (the context, reference and response in our case,) multiplicative systems take the form

$$\sigma_m(c, r, \hat{r}) = \hat{r}^T f(c, r) \tag{3}$$

whereas an additive system takes the form

$$\sigma_a(c, r, \hat{r}) = \hat{r} + f(c, r) \tag{4}$$

Note that the function f can be selected such that the system is uniformly multiplicative or additive for all the components, or vice-versa, wherein two of the components interact additive and the remaining have multiplicative interactions. For ADEM this function f can be written as

$$f(c, r) = A^T c + B^T r \tag{5}$$

which is a linear additive function. Given a set of embedding vectors $V = [v_1, v_2, \ldots]$, the two metrics from Chandrahas et al. (2018), are defined as

- Alignment to Mean (ATM)

  $$\text{ATM}(v, \bar{V}) = v^T \bar{V}$$

  where $\bar{V}$ is the mean of all the vectors in V.

- Conicity of the set of vectors is the mean of the ATMs for all the vectors in the set

  $$\text{Conicity}(V) = \frac{1}{|V|} \sum_{x \in V} \text{ATM}(x^T \bar{V})$$

Chandrahas et al. (2018) empirically show that multiplicative systems lead to embedding spaces with high conicity and hence all the responses get closely huddled around the mean as shown in Fig. 1.

The principal finding from our analysis so far points towards the fact that the scores generated by ADEM are all tightly clustered around the mean value and hence the discriminative power of the system is limited at best. Taking a deeper look at the ADEM cost function we can write

$$\begin{aligned} score(c, r, \hat{r}) &= (c^T M \hat{r} + r^T N \hat{r} - \alpha)/\beta & (6) \\ &= a^T \hat{r} - b & (7) \end{aligned}$$

Response to be evaluated	mean	SD	%1 SD
Reference response	2.75	0.34	71.65
Punctuation removed	2.85	0.31	71.65
NLTK stopwords removed	2.69	0.33	70.60
25 common stopwords removed	2.80	0.24	69.08
[pro]nouns and verbs only	2.80	0.36	68.96
Named entities removed	2.74	0.35	70.60
Replace words with synonyms	2.83	0.32	70.36
Jumble words in the sentence	2.73	0.33	72
Reverse the response	2.75	0.33	68.84
Retain only nouns	2.73	0.39	68.26
Repeat words in the response	2.70	0.36	71.41
Generic and Irrelevant responses:			
I'm sorry, can you repeat?	2.65	0.34	69.43
I will do	2.69	0.34	70
fantastic! how are you?	3.18	0.4	69.4

Table 4: ADEM scores on simple dataset variants. The last column indicates the percentage of scores within one standard deviation of the mean score.

where $a = (c^T M + r^T N)/\beta$ and $b = \alpha/\beta$. Without loss of generality, assume that there are two model responses $\hat{r}_1$ and $\hat{r}_2$ for the same context c and reference r. Since the model is linear and all other variables are fixed, for a new (valid) response $\hat{r}_n$ which lies on $\hat{r}_2 - \hat{r}_1$ (or more generically a linear combination of $\hat{r}_1$ and $\hat{r}_2$), we can write the new score as

$$\begin{aligned} s_n &= a^T \hat{r}_n - b & (8) \\ &= a^T \left(\hat{r}_1 + \Delta \frac{\hat{r}_2 - \hat{r}_1}{M} \right) - b & (9) \\ &= \left(s_1 + \frac{\Delta}{M}(s_2 - s_1) \right) - b & (10) \end{aligned}$$

where Δ is a small step in the direction of $\hat{r}_2 - \hat{r}_1$, M is the magnitude of the vector $\|\hat{r}_2 - \hat{r}_1\|$. Assuming $s_2 > s_1$, the score difference $(s_n - s_1)$ can be maximized (leading to diverse scores for close vectors), by taking $M \to 0$ which leads to $\hat{r}_1 \to \hat{r}_2$ which leads to mapping all the response vectors very close to each other leading to high conicity and low vector spread.

Our analysis for ADEM points to a similar artifact. ADEM's response embeddings of $200k$ twitter responses has a conicity value of 0.59. Further we note that the score is a linear formulation [see Eq. 1]. This means that for a given context and reference response, there is no way to score two valid responses high by circumventing any bad embeddings that lie in the subspace spanned by these two responses.

Analysing Robustness of ADEM to Adversarial Attacks

For the following section of human constructed attacks, we run the experiments on the Microsoft Research Social Media Conversation Corpus (Sordoni et al. 2015). This corpus contains a curated list of 3 turn twitter conversations, all of which received an average score of 4 or higher for good response quality by 3 crowd-sourced annotators apiece. Human scoring of the dialogues was performed on a 5-point

scale similar to ADEM's automated scoring on a scale [1-5] which makes it easier for comparison.
We construct some simple test cases by modifying the reference response and summarize the findings in Table 4. This analysis on ADEM shows:

- multiple adversarial scenarios where the evaluator is easily fooled or does the opposite of what is expected
- the ADEM score in most cases is a conservative average value of around 2.75

Although these experiments are reported on ADEM model, they can be used to analyze any other dialog evaluator model. Table 4 and Table 5 show the results of the experiments with various dataset modifications of different flavors explained below:

- **Evaluation of the Ground Truth Response** to estimate the reliability of the evaluator. We evaluate the scores on the responses in the training set of MSR Social Media Conversation corpus, which are already rated high by human annotators. This first experiment revealed that the ADEM scores are concentrated around the average value of 2.75 which seems like a conservative middle value on a scale of [1-5]. The standard deviation of the scores is 0.34 with 71.65% of the scores within the first standard deviation.
- **Slight Modifications** such as removing punctuation and stopwords should probably lower the score only a little for a given response.

  Removing Punctuation In this modification of the dataset, the test responses are stripped of all punctuation marks. We note that a human's score of such a response would still be just about the same. We conclude this from an human evaluation of a sample of 150 such responses by crowd-sourced annotators resulting in an average score of 4.83. However, the correlation of the scores by ADEM on responses with and without this modification is not that strong.

  Removing Stopwords Removing stop words should not affect the score much if one is lenient with the grammar. The standard English stop words from NLTK (Natural Language ToolKit) library of python were removed in this variant (NLTK stopwords removed). Another variant where only the top 25 most common stop words were removed in also listed in Table 4 as 25 common stopwords removed. The list of the 25 is as follows:

  a, an, and, are, as, at, be, by, for, from,
  has, he, in, is, it, its, of, on, that, the,
  to, was, were, will, with

  A peculiarity that can be observed is the better correlation of scores when more stopwords are removed than when just the 25 most common ones are filtered out. While this could be explained as more stopwords eliminated leading to reduction in noise and unnecessary dimensions, it is questionable whether human scoring would follow the same pattern. The human evaluation on 150 samples of each of these types decrements average scores to 4.6 and 4.2 respectively when only 25 stopwords are removed and all the NLTK stopwords are removed. However, Table 5 shows that the score is better than that of the original response in over 60% cases when either punctuation or the 25 most common stopwords are removed.

Variant	Pearson	Spearman	Better score
Punctuation removed	0.55	0.5	64.41%
NLTK stopwords removed	0.78	0.76	37.92%
25 common stopwords removed	0.6	0.57	60.33%
[pro]nouns and verbs	0.52	0.49	56.71%
Named entities removed	0.98	0.97	11.2%
Replace words with synonyms	0.79	0.75	68.03%
Jumble words in the sentence	0.68	0.64	47.02%
Reverse the response	0.52	0.49	48.66%
Retain only nouns	0.29	0.26	50.64%
Repeat words in the response	0.91	0.90	37.57%
"fantastic! how are you?"	0.34	0.32	86.93%

Table 5: Correlation of ADEM scores on different variants of the response with the ADEM scores on original (reference) response. p-values in all these cases are < 0.001. The last column indicates the percentage of times the concerned variant received a better score than original

- **Simplifying the Response:**

  Retaining Only the Nouns, Pronouns and Verbs in the Response Our intention was to check if a simpler version of the response can be created by removing adjectives, adverbs, etc and retaining just the nouns, pronouns and verbs. We performed a human evaluation of 150 dialogue samples to check if the core idea of the modified responses would still be the same in most cases. We find the human scores drop to an average of 2.78. We hence observe that the hypothesis is not true for all responses and do not pose expectations on correlation of scores for this variant.

  Removing Named Entities An even milder modification is to just remove the named entities which we identify using the POS (Part of Speech) tagger functions and Named Entity Recognition modules of the NLTK library. The score isn't expected to be affected much by such a modification. This is also reflected by the strong correlation in the scores of this variant with the original response.
- **Replacing Words by Synonyms:** This variant contains most of the words replaced by their synonyms (excluding stop words and named entities) using WordNet from NLTK. Since the syntax, semantics and the meaning of the response are unaltered, the score should ideally be [almost] the same and correlation should be high. ADEM scores exhibit a decent correlation between the response and its variant with synonyms-replaced.
- **Perturbing the Response** by reversing the sentence or jumbling the words, randomly repeating words in the response create unnatural responses.

  Jumble the Sentence If the response words appear in jumbled order, low scores are expected. Also it would

be an interesting insight into the working of an evaluator model to check whether the scores on this variant correlate with the original response. From the correlation values in Table 5 we observe that ADEM does not take into account the syntax and semantics of a response while scoring it.

Reverse the Sentence This can be considered as a special case of the previous variant. The word order is the exact reverse of the original.

Retain Only Nouns in the Response Using the POS tagger functionality of NLTK library, the utterance was modified to contain only the nouns (proper and common nouns). This should ideally obtain low scores from a dialogue evaluator. This variant indeed elicits the weakest correlation with the original response which seems logical as the meaning of the response is lost.

Repeat Words in the Response An unnatural variant of a response/utterance is created in this case by randomly choosing $m/2$ words to be repeated in the utterance, where m is the length of the utterance. An evaluator should be able to distinguish when a repetition is supposed to be penalized and when it is natural. Whereas, it can be observed that the scores on the response with repetitions are strongly correlated to the original responses.

- **Generic and Irrelevant Responses** Table 4 also shows the scores when a generic response is used, such as "I'm sorry, can you repeat?" which would be applicable as a suitable response in most cases. Also a dialogue evaluation metric should not possess a loophole where a dialogue generation system can learn to pick up a certain utterance in the vector space with which it can consistently manage to get high scores from the evaluator. The response "Fantastic! How are you?" seems to be one such case with ADEM, where it is given relatively higher scores in any context. It was found to be scored higher than the reference response by ADEM in 86.93% of the cases.

Although these are very simple cases, the goal towards a good dialogue evaluation model would be realized only with the ability to handle such cases.

Whitebox Attack on ADEM

While a lot of work has been done around fooling convolutional networks using gradient descent, the methods of adding noise to fool an NLP model is not straightforward due to the discrete nature of text. Works towards this goal hence try various ways of perturbing the data or producing adversarial/unfavorable examples. Targeting the Question Answering models, Jia and Liang (2017) insert adversarial sentences in the SQuAD dataset (Rajpurkar et al. 2016), which do not affect the correct answer or mislead humans. However, they report a drop in the average accuracy of sixteen published models from 75% to 36% F1 score. Li et al. (2016) propose to interpret the importance of various aspects of a neural network by analyzing the effects of erasing various parts of the representations such as input word-vector dimensions, intermediate hidden units and input words. Liang et al. (2018) fool text classification by either adding, removing or modifying the inputs. Zhao et al. (2018) add perturbations in the continuous space rather than at the discrete input level, using autoencoders to map the discrete text into continuous codes.

Type	mean	std dev	max	min
Original responses	2.75	0.34	4.19	1.44
Brute force search	3.93	0.24	4.7	3.39
Annoy index search	3.62	0.29	4.3	3.24

Table 6: Statistics of the ADEM scores on various responses

Context	Response	ADEM	human
<first_speaker> Hi, how are you doing? <second_speaker> not too bad, getting over a cold	colds are no good ,hope it clears away soon.	2.9	5
	<first_speaker> sure, you can get an autographed one right here: <URL>	3.9	1
	<first_speaker> mate, get the app, you can watch while running,	3.79	2.5
<first_speaker>Awesome ! Did you graduate high school this year ? <second_speaker> Nope 2 more years !	<first_speaker> Wow ! Going to be a junior . You youngster !	2.4	5
	<first_speaker> that is what sundays are for !	3.6	1
	<first_speaker> better safe than sorry !	3.56	1

Table 7: Examples with ADEM and human scores

In the context of neural network based dialogue evaluators, we propose a white box attack to game the model for high scores (or low scores) given a context and a reference response. We employ guided backpropagation (Springenberg et al. 2014) from the score function back to the response sentence embedding level of ADEM model. This directs changes in the response embedding towards producing the desired score. Employing this method, we arrive at an embedding that produces a score between $4.6 - 4.9$. We now need the sentence/response that translates to this computed embedding that can fetch the desired score. Here, we settle for an approximation of the desired score by finding the sentence closest to the desired sentence embedding computed. For this, we use a database of sentence embeddings constructed on 470k Twitter responses crawled using the Twitter IDs given by Hori and Hori (2017) for customer service conversations. The embeddings are indexed using Annoy index (Li et al. 2016b), which is designed to identify approximate nearest neighbors in a multidimensional space. We query the corpus using this index with our response embedding to

Context	Reference response	Response to evaluate	ADEM
do you want to watch a movie today ?	will it rain tonight ?	will it rain tonight ?	2.22
		yes but will it rain tonight ?	2.52
		sure , avengers infinity war is out	2.37
		i did not finish my assignment	2.49
		tonight rain it will ?	2.37
		? tonight rain it will	**2.7**

Table 8: ADEM scores on an example context with various responses

obtain its [approximate] nearest neighbor embeddings. We gather 400 of these nearest neighbors and find the best score on these by ADEM. The scores increase on average by 0.87 to form an average score of 3.62. To understand the upper limit or the potential for improvement in the score when using this database, we also employ brute force search on the entire embeddings database to find the best score given by ADEM. In other words, we compute the scores on all the response embeddings in our database and pick the highest. The mean score in this case is 3.93 with an average increment of 1.18. These results are tabulated in Table 6

Human Evaluation

We check the relevance of the responses that were scored high by ADEM for the given context. The human evaluation is made by in-house annotators using 2 approaches:

- scoring the top response, as selected by ADEM, on a scale [1-5] similar to ADEM's scoring scheme
- re-ranking ADEM's top 5 scored responses. The dialogues are presented in the order of ADEM scores (high-low). The annotators assign their rank/preference of each response for the given the context.

We find that the average human rating on the responses scored highest by ADEM is 1.9 for 250 samples. The re-ranking experiment could not be followed as intended as most of the responses were bad leading to no preferences in the responses by the annotators in most cases.
Table 7 shows some examples of human and ADEM scores on original responses and the responses fetched by brute force or annoy index search.

Conclusion and Future Directions

We summarize our findings with an example (Table 8) to depict the current state of dialogue evaluators and highlight the scope and need for improvement. From our study, we identify the following requirements towards the goal of an ideal evaluator:

- handle scoring diverse valid responses high
- sensitivity to grammar and relevance of the response
- not be heavily influenced by the reference response
- robust against fooling attacks

In this work we develop a systematic approach to attack a dialogue evaluation system. The attacks point to multiple simple modifications to responses which will have very low correlation with human responses, but still garner high scores from ADEM. Our experiments can be used by the research community to guide a system similar to ADEM towards better performance thereby leading to higher human correlation with the response scores. We believe, from our analysis of ADEM model, that a non-linear scoring model might be better suited for the task of dialogue evaluation. We sincerely hope that the attacks mentioned in this work become the guiding principles for the design of the dialogue evaluation module of the future.

Acknowledgements

We thank NVIDIA and Robert Bosch Center for Data Sciences and Artificial Intelligence (RBC-DSAI), IIT Madras for compute resources. We also thank Madhuri, Govind, Samprit, Nikhil and Baladitya for helping us with human evaluations.

References

Banerjee, S., and Lavie, A. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, 65–72. Association for Computational Linguistics.

Callison-Burch, C.; Osborne, M.; and Koehn, P. 2006. Re-evaluation the role of bleu in machine translation research. In *EACL*. The Association for Computer Linguistics.

Callison-Burch, C. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *EMNLP*, 286–295. ACL.

Chandrahas; Sharma, A.; and Talukdar, P. 2018. Towards understanding the geometry of knowledge graph embeddings. In *ACL*. ACL.

Galley, M.; Brockett, C.; Sordoni, A.; Ji, Y.; Auli, M.; Quirk, C.; Mitchell, M.; Gao, J.; and Dolan, B. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. In *ACL (2)*, 445–450. The Association for Computer Linguistics.

Gao, J.; Galley, M.; and Li, L. 2018. Neural approaches to conversational AI. *CoRR* abs/1809.08267.

Hori, C., and Hori, T. 2017. End-to-end conversation modeling track in dstc6. *arXiv:1706.07440*.

Jia, R., and Liang, P. 2017. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*, 2021–2031. Association for Computational Linguistics.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. Association for Computational Linguistics.

Li, W.; Zhang, Y.; Sun, Y.; Wang, W.; Zhang, W.; and Lin, X. 2016b. Approximate nearest neighbor search on high dimensional data - experiments, analyses, and improvement (v1.0). *CoRR* abs/1610.02455.

Li, J.; Monroe, W.; Shi, T.; Ritter, A.; and Jurafsky, D. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Li, J.; Monroe, W.; and Jurafsky, D. 2016. Understanding neural networks through representation erasure. *CoRR* abs/1612.08220.

Liang, B.; Li, H.; Su, M.; Bian, P.; Li, X.; and Shi, W. 2018. Deep text classification can be fooled. In *IJCAI*, 4208–4215. ijcai.org.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Liu, C.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, 2122–2132. The Association for Computational Linguistics.

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL Conference*, 285–294. The Association for Computer Linguistics.

Lowe, R.; Noseworthy, M.; Serban, I. V.; Angelard-Gontier, N.; Bengio, Y.; and Pineau, J. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *ACL (1)*, 1116–1126. Association for Computational Linguistics.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, 3111–3119.

Nema, P., and Khapra, M. M. 2018. Towards a better metric for evaluating question generation systems. *CoRR* abs/1808.10192.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318. ACL.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543. ACL.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, 2383–2392. The Association for Computational Linguistics.

Ritter, A.; Cherry, C.; and Dolan, B. 2010. Unsupervised modeling of twitter conversations. In *HLT-NAACL*, 172–180. The Association for Computational Linguistics.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-driven response generation in social media. In *EMNLP*, 583–593. ACL.

Saha, A.; Khapra, M. M.; and Sankaranarayanan, K. 2018. Towards building large scale multimodal domain-aware conversation systems. In *AAAI*. AAAI Press.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.; Gao, J.; and Dolan, B. 2015. A neural network approach to context-sensitive generation of conversational responses. In *HLT-NAACL*, 196–205. The Association for Computational Linguistics.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. A. 2014. Striving for simplicity: The all convolutional net. *CoRR* abs/1412.6806.

Vinyals, O., and Le, Q. V. 2015. A neural conversational model. *CoRR* abs/1506.05869.

Wen, T.; Gasic, M.; Mrksic, N.; Su, P.; Vandyke, D.; and Young, S. J. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *EMNLP*, 1711–1721. The Association for Computational Linguistics.

Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2852–2858.

Zhao, Z.; Dua, D.; and Singh, S. 2018. Generating natural adversarial examples. *International Conference on Learning Representations*.