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

Chia-Wei Liu[1]*, **Ryan Lowe**[1]*, **Iulian V. Serban**[2]*, **Michael Noseworthy**[1]*,
Laurent Charlin[1], **Joelle Pineau**[1]
[1] School of Computer Science, McGill University
{chia-wei.liu,ryan.lowe,michael.noseworthy}@mail.mcgill.ca
{lcharlin, jpineau}@cs.mcgill.ca
[2] DIRO, Université de Montréal
iulian.vlad.serban@umontreal.ca

Abstract

We investigate evaluation metrics for dialogue response generation systems where supervised labels, such as task completion, are not available. Recent works in response generation have adopted metrics from machine translation to compare a model's generated response to a single target response. We show that these metrics correlate very weakly with human judgements in the non-technical Twitter domain, and not at all in the technical Ubuntu domain. We provide quantitative and qualitative results highlighting specific weaknesses in existing metrics, and provide recommendations for future development of better automatic evaluation metrics for dialogue systems.

1 Introduction

An important aspect of dialogue response generation systems, which are trained to produce a reasonable utterance given a conversational context, is how to evaluate the quality of the generated response. Typically, evaluation is done using human-generated supervised signals, such as a task completion test or a user satisfaction score (Walker et al., 1997; Möller et al., 2006; Kamm, 1995), which are relevant when the dialogue is task-focused. We call models optimized for such supervised objectives *supervised dialogue models*, while those that do not are *unsupervised dialogue models*.

This paper focuses on unsupervised dialogue response generation models, such as chatbots. These models are receiving increased attention, particularly using end-to-end training with neural networks (Serban et al., 2016; Sordoni et al., 2015; Vinyals and Le, 2015). This avoids the need to collect supervised labels on a large scale, which can be prohibitively expensive. However, automatically evaluating the quality of these models remains an open question. Automatic evaluation metrics would help accelerate the deployment of unsupervised response generation systems.

Faced with similar challenges, other natural language tasks have successfully developed automatic evaluation metrics. For example, BLEU (Papineni et al., 2002a) and METEOR (Banerjee and Lavie, 2005) are now standard for evaluating machine translation models, and ROUGE (Lin, 2004) is often used for automatic summarization. These metrics have recently been adopted by dialogue researchers (Ritter et al., 2011; Sordoni et al., 2015; Li et al., 2015; Galley et al., 2015b; Wen et al., 2015; Li et al., 2016). However these metrics assume that valid responses have significant word overlap with the ground truth responses. This is a strong assumption for dialogue systems, where there is significant diversity in the space of valid responses to a given context. This is illustrated in Table 1, where two reasonable responses are proposed to the context, but these responses do not share any words in common and do not have the same semantic meaning.

In this paper, we investigate the correlation between the scores from several automatic evaluation metrics and human judgements of dialogue response quality, for a variety of response generation models. We consider both statistical *word-overlap similar-*

* Denotes equal contribution.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2122–2132,
Austin, Texas, November 1-5, 2016.

Context of Conversation
Speaker A: Hey John, what do you want to do tonight?
Speaker B: Why don't we go see a movie?
Ground-Truth Response
Nah, I hate that stuff, let's do something active.
Model Response
Oh sure! Heard the film about Turing is out!

Table 1: Example showing the intrinsic diversity of valid responses in a dialogue. The (reasonable) model response would receive a BLEU score of 0.

ity metrics such as BLEU, METEOR, and ROUGE, and *word embedding metrics* derived from word embedding models such as Word2Vec (Mikolov et al., 2013). We find that all metrics show either weak or no correlation with human judgements, despite the fact that word overlap metrics have been used extensively in the literature for evaluating dialogue response models (see above, and Lasguido et al. (2014)). In particular, we show that these metrics have only a small positive correlation on the chitchat oriented Twitter dataset, and no correlation at all on the technical Ubuntu Dialogue Corpus. For the word embedding metrics, we show that this is true even though all metrics are able to significantly distinguish between baseline and state-of-the-art models across multiple datasets. We further highlight the shortcomings of these metrics using: a) a statistical analysis of our survey's results; b) a qualitative analysis of examples from our data; and c) an exploration of the sensitivity of the metrics.

Our results indicate that a shift must be made in the research community away from these metrics, and highlight the need for a new metric that correlates more strongly with human judgement.

2 Related Work

We focus on metrics that are *model-independent*, i.e. where the model generating the response does not also evaluate its quality; thus, we do not consider word perplexity, although it has been used to evaluate unsupervised dialogue models (Serban et al., 2015). This is because it is not computed on a per-response basis, and cannot be computed for retrieval models. Further, we only consider metrics that can be used to evaluate proposed responses against ground-truth responses, so we do not consider retrieval-based metrics such as recall, which has been used to evaluate dialogue models (Schatzmann et al., 2005; Lowe et al., 2015). We also do not consider evaluation methods for supervised evaluation methods.[1]

Several recent works on unsupervised dialogue systems adopt the BLEU score for evaluation. Ritter et al. (2011) formulate the unsupervised learning problem as one of translating a context into a candidate response. They use a statistical machine translation (SMT) model to generate responses to various contexts using Twitter data, and show that it outperforms information retrieval baselines according to both BLEU and human evaluations. Sordoni et al. (2015) extend this idea using a recurrent language model to generate responses in a context-sensitive manner. They also evaluate using BLEU, however they produce multiple ground truth responses by retrieving 15 responses from elsewhere in the corpus, using a simple bag-of-words model. Li et al. (2015) evaluate their proposed diversity-promoting objective function for neural network models using BLEU score with only a single ground truth response. A modified version of BLEU, deltaBLEU (Galley et al., 2015b), which takes into account several human-evaluated ground truth responses, is shown to have a weak to moderate correlation to human judgements using Twitter dialogues. However, such human annotation is often infeasible to obtain in practice. Galley et al. (2015b) also show that, even with several ground truth responses available, the standard BLEU metric does not correlate strongly with human judgements.

There has been significant previous work that evaluates how well automatic metrics correlate with human judgements in in both machine translation (Callison-Burch et al., 2010; Callison-Burch et al., 2011; Bojar et al., 2014; Graham et al., 2015) and natural language generation (NLG) (Stent et al., 2005; Cahill, 2009; Reiter and Belz, 2009; Espinosa et al., 2010). There has also been work criticizing the usefulness of BLEU in particular for machine translation (Callison-Burch et al., 2006). While many of the criticisms in these works apply to dialogue generation, we note that generating dialogue responses conditioned on the conversational

[1] Evaluation methods in the supervised setting have been well studied, see (Walker et al., 1997; Möller et al., 2006; Jokinen and McTear, 2009).

context is in fact a more difficult problem. This is because most of the difficulty in automatically evaluating language generation models lies in the large set of correct answers. Dialogue response generation given solely the context intuitively has a higher diversity (or entropy) than translation given text in a source language, or surface realization given some intermediate form (Artstein et al., 2009).

3 Evaluation Metrics

Given a dialogue context and a proposed response, our goal is to automatically evaluate how appropriate the proposed response is to the conversation. We focus on metrics that compare it to the ground truth response of the conversation. In particular, we investigate two approaches: word based similarity metrics and word-embedding based similarity metrics.

3.1 Word Overlap-based Metrics

We first consider metrics that evaluate the amount of *word-overlap* between the proposed response and the ground-truth response. We examine the BLEU and METEOR scores that have been used for machine translation, and the ROUGE score that has been used for automatic summarization. While these metrics have been shown to correlate with human judgements in their target domains (Papineni et al., 2002a; Lin, 2004), they have not been thoroughly investigated for dialogue systems.[2]

We denote the ground truth response as r (thus we assume that there is a single candidate ground truth response), and the proposed response as $\hat{r}$. The j'th token in the ground truth response r is denoted by w_j, with $\hat{w}_j$ denoting the j'th token in the proposed response $\hat{r}$.

BLEU. BLEU (Papineni et al., 2002a) 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 (we assume that there is a single candidate ground truth response per context):

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

where k indexes all possible n-grams of length n and $h(k, r)$ is the number of n-grams k in r.[3] 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 uses $N = 4$. Modern versions of BLEU also use sentence-level smoothing, as 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.

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.

[2] To the best of our knowledge, only BLEU has been evaluated in the dialogue system setting quantitatively by Galley et al. (2015a) on the Twitter domain. However, they carried out their experiments in a very different setting with multiple ground truth responses, which are rarely available in practice, and without providing any qualitative analysis of their results.

[3] 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.

3.2 Embedding-based Metrics

An alternative to using word-overlap based metrics is to consider the meaning of each word as defined by a *word embedding*, which assigns a vector to each word. Methods such as Word2Vec (Mikolov et al., 2013) calculate these embeddings using distributional semantics; that is, they approximate the meaning of a word by considering how often it co-occurs with other words in the corpus.[4] These embedding-based metrics usually approximate sentence-level embeddings using some heuristic to combine the vectors of the individual words in the sentence. The sentence-level embeddings between the candidate and target response are compared using a measure such as cosine distance.

Greedy Matching. Greedy matching is the one embedding-based metric that does not compute sentence-level embeddings. Instead, given two sequences r and $\hat{r}$, each token $w \in r$ is greedily matched with a token $\hat{w} \in \hat{r}$ based on the cosine similarity of their word embeddings (e_w), and the total score is then averaged across all words:

$$G(r, \hat{r}) = \frac{\sum_{w \in r;} \max_{\hat{w} \in \hat{r}} cos\_sim(e_w, e_{\hat{w}})}{|r|}$$

$$GM(r, \hat{r}) = \frac{G(r, \hat{r}) + G(\hat{r}, r)}{2}$$

This formula is asymmetric, thus we must average the greedy matching scores G in each direction. This was originally introduced for intelligent tutoring systems (Rus and Lintean, 2012). The greedy approach favours responses with key words that are semantically similar to those in the ground truth response.

Embedding Average. The embedding average metric calculates sentence-level embeddings using additive composition, a method for computing the meanings of phrases by averaging the vector representations of their constituent words (Foltz et al., 1998; Landauer and Dumais, 1997; Mitchell and Lapata, 2008). This method has been widely used in other domains, for example in textual similarity tasks (Wieting et al., 2015). The embedding average, $\bar{e}$, is defined as the mean of the word embeddings of each token in a sentence r:

$$\bar{e}_r = \frac{\sum_{w \in r} e_w}{|\sum_{w' \in r} e_{w'}|}.$$

To compare a ground truth response r and retrieved response $\hat{r}$, we compute the cosine similarity between their respective sentence level embeddings: EA $:= \cos(\bar{e}_r, \bar{e}_{\hat{r}})$.

Vector Extrema. Another way to calculate sentence-level embeddings is using vector extrema (Forgues et al., 2014). For each dimension of the word vectors, take the most extreme value amongst *all word vectors in the sentence*, and use that value in the sentence-level embedding:

$$e_{rd} = \begin{cases} \max_{w \in r} e_{wd} & \text{if } e_{wd} > |\min_{w' \in r} e_{w'd}| \\ \min_{w \in r} e_{wd} & \text{otherwise} \end{cases}$$

where d indexes the dimensions of a vector; e_{wd} is the d'th dimensions of e_w (w's embedding). The min in this equation refers to the selection of the largest negative value, if it has a greater magnitude than the largest positive value.

Similarity between response vectors is again computed using cosine distance. Intuitively, this approach prioritizes informative words over common ones; words that appear in similar contexts will be close together in the vector space. Thus, common words are pulled towards the origin because they occur in various contexts, while words carrying important semantic information will lie further away. By taking the extrema along each dimension, we are thus more likely to ignore common words.

4 Dialogue Response Generation Models

In order to determine the correlation between automatic metrics and human judgements of response quality, we obtain response from a diverse range of response generation models in the recent literature, including both *retrieval* and *generative* models.

4.1 Retrieval Models

Ranking or retrieval models for dialogue systems are typically evaluated based on whether they can retrieve the correct response from a corpus of predefined responses, which includes the ground truth

[4] To maintain statistical independence between the task and each performance metric, it is important that the word embeddings used are trained on corpora which do not overlap with the task corpus.

	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.

response to the conversation (Schatzmann et al., 2005). Such systems can be evaluated using recall or precision metrics. However, when deployed in a real setting these models will not have access to the correct response given an unseen conversation. Thus, in the results presented below we *remove* one occurrence of the ground-truth response from the corpus and ask the model to retrieve the most appropriate response from the remaining utterances. Note that this does not mean the correct response will not appear in the corpus at all; in particular, if there exists another context in the dataset with an identical ground-truth response, this will be available for selection by the model.

We then evaluate each model by comparing the retrieved response to the ground truth response of the conversation. This closely imitates real-life deployment of these models, as it tests the ability of the model to generalize to unseen contexts.

TF-IDF. We consider a simple Term Frequency - Inverse Document Frequency (TF-IDF) retrieval model (Lowe et al., 2015). TF-IDF is a statistic that intends to capture how important a given word is to some document, which is calculated as: $\text{tfidf}(w, c, C) = f(w, c) \times \log \frac{N}{|\{c \in C : w \in c\}|}$, where C is the set of all contexts in the corpus, $f(w, c)$ indicates the number of times word w appeared in context c, N is the total number of dialogues, and the denominator represents the number of dialogues in which the word w appears.

In order to apply TF-IDF as a retrieval model for dialogue, we first compute the TF-IDF vectors for each context and response in the corpus. We then return the response with the largest cosine similarity in the corpus, either between the input context and corpus contexts (C-TFIDF), or between the input context and corpus responses (R-TFIDF).

Dual Encoder. Next we consider the recurrent neural network (RNN) based architecture called the Dual Encoder (DE) model (Lowe et al., 2015). The DE model consists of two RNNs which respectively compute the vector representation of an input context and response, $c, r \in \mathbb{R}^n$. The model then calculates the probability that the given response is the ground truth response given the context, by taking a weighted dot product: $p(r \text{ is correct} | c, r, M) = \sigma(c^T M r + b)$ where M is a matrix of learned parameters and b is a bias. The model is trained using negative sampling to minimize the cross-entropy error of all (context, response) pairs. To our knowledge, our application of neural network models to large-scale retrieval in dialogue systems is novel.

4.2 Generative Models

In addition to retrieval models, we also consider generative models. In this context, we refer to a model as generative if it is able to generate entirely new sentences that are unseen in the training set.

LSTM language model. The baseline model is an LSTM language model (Hochreiter and Schmidhuber, 1997) trained to predict the next word in the (context, response) pair. During test time, the model is given a context, encodes it with the LSTM and generates a response using a greedy beam search procedure (Graves, 2013).

HRED. Finally we consider the Hierarchical Recurrent Encoder-Decoder (HRED) (Serban et al., 2015). In the traditional Encoder-Decoder framework, all utterances in the context are concatenated together before encoding. Thus, information from previous utterances is far outweighed by the most recent utterance. The HRED model uses a *hierarchy* of encoders; each utterance in the context passes through an 'utterance-level' encoder, and the

	Twitter				Ubuntu			
Metric	**Spearman**	p-value	**Pearson**	p-value	**Spearman**	p-value	**Pearson**	p-value
Greedy	0.2119	0.034	0.1994	0.047	0.05276	0.6	0.02049	0.84
Average	0.2259	0.024	0.1971	0.049	-0.1387	0.17	-0.1631	0.10
Extrema	0.2103	0.036	0.1842	0.067	0.09243	0.36	-0.002903	0.98
METEOR	0.1887	0.06	0.1927	0.055	0.06314	0.53	0.1419	0.16
BLEU-1	0.1665	0.098	0.1288	0.2	-0.02552	0.8	0.01929	0.85
BLEU-2	0.3576	< 0.01	0.3874	< 0.01	0.03819	0.71	0.0586	0.56
BLEU-3	0.3423	< 0.01	0.1443	0.15	0.0878	0.38	0.1116	0.27
BLEU-4	0.3417	< 0.01	0.1392	0.17	0.1218	0.23	0.1132	0.26
ROUGE	0.1235	0.22	0.09714	0.34	0.05405	0.5933	0.06401	0.53
Human	0.9476	< 0.01	1.0	0.0	0.9550	< 0.01	1.0	0.0

Table 3: Correlation between each metric and human judgements for each response. Correlations shown in the human row result from randomly dividing human judges into two groups.

	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.

	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

Table 5: Effect of differences in response length for the Twitter dataset, Δw = absolute difference in #words between a ground truth response and proposed response

output of these encoders is passed through another 'context-level' encoder, which enables the handling of longer-term dependencies.

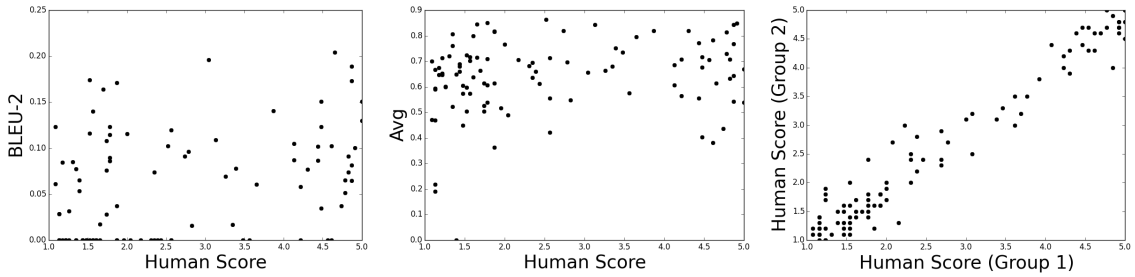
4.3 Conclusions from an Incomplete Analysis

When evaluation metrics are not explicitly correlated to human judgement, it is possible to draw misleading conclusions by examining how the metrics rate different models. To illustrate this point, we compare the performance of selected models according to the embedding metrics on two different domains: the Ubuntu Dialogue Corpus (Lowe et al., 2015), which contains technical vocabulary and where conversations are often oriented towards solving a particular problem, and a non-technical Twitter corpus collected following the procedure of Ritter et al. (2010). We consider these two datasets since they cover contrasting dialogue domains, i.e. technical help vs casual chit-chat, and because they are amongst the largest publicly available corpora, making them good candidates for building data-driven dialogue systems.

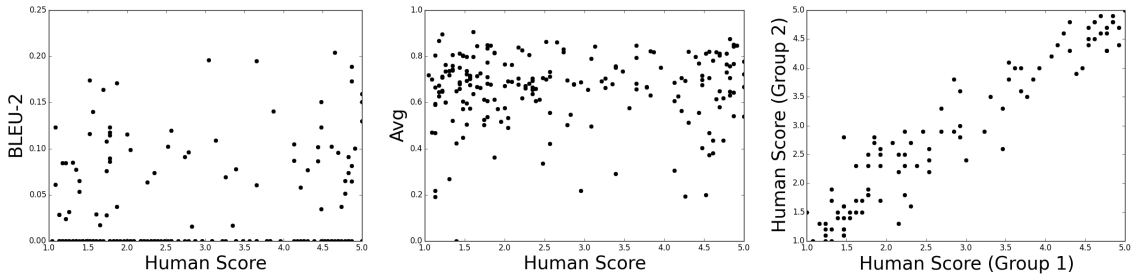
Results on the proposed embedding metrics are shown in Table 2. For the retrieval models, we observe that the DE model significantly outperforms both TFIDF baselines on all metrics across both datasets. Further, the HRED model significantly outperforms the basic LSTM generative model in both domains, and appears to be of similar strength as the DE model. Based on these results, one might be tempted to conclude that there is some information being captured by these metrics, that significantly differentiates models of different quality. However, as we show in the next section, the embedding-based metrics correlate only weakly with human judgements on the Twitter corpus, and not at all on the Ubuntu Dialogue Corpus. This demonstrates that metrics that have not been specifically correlated with human judgements on a new task should not be used to evaluate that task.

5 Human Correlation Analysis

Data Collection. We conducted a human survey to determine the correlation between human judgements on the quality of responses, and the score assigned by each metric. We aimed to follow the procedure for the evaluation of BLEU (Papineni et al.,



(a) Twitter



(b) Ubuntu

Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

2002a). 25 volunteers from the Computer Science department at the author's institution were given a context and one proposed response, and were asked to judge the response quality on a scale of 1 to 5.[5]; a 1 indicates that the response is not appropriate or sensible given the context, and a 5 indicates that the response is very reasonable. Out of the 25 respondents, 23 had Cohen's kappa scores $\kappa > 0.2$ w.r.t. the other respondents, which is a standard measure for inter-rater agreement (Cohen, 1968). The 2 respondents with $\kappa < 0.2$, indicating slight agreement, were excluded from the analysis below. The median κ score was approximately 0.55, roughly indicating moderate to strong annotator agreement.

Each volunteer was given 100 questions per dataset. These questions correspond to 20 unique contexts, with 5 different responses: one utterance randomly drawn from elsewhere in the test set, the response selected from each of the TF-IDF, DE, and HRED models, and a response written by a human annotator. These were chosen as they cover the range of qualities almost uniformly (see Figure 1).

Survey Results. We present correlation results between the human judgements and each metric in Table 3. We compute the Pearson correlation, which estimates linear correlation, and Spearman correlation, which estimates any monotonic correlation.

The first observation is that in both domains the BLEU-4 score, which has previously been used to evaluate unsupervised dialogue systems, shows very weak if any correlation with human judgement. In fact we found that the BLEU-3 and BLEU-4 scores were near-zero for a majority of response pairs; for BLEU-4, only four examples had a score $> 10^{-9}$. Despite this, they still correlate with human judgements on the Twitter Corpus at a rate similar to BLEU-2. This is because of the smoothing constant, which gives a tiny weight to unigrams and bigrams despite the absence of higher-order n-grams. BLEU-3 and BLEU-4 behave as a scaled, noisy version of BLEU-2; thus, if one is to evaluate dialogue

[5]Studies asking humans to evaluate text often rate different aspects separately, such as 'adequacy', 'fluency' and 'informativeness' of the text (Hovy, 1999; Papineni et al., 2002b) Our evaluation focuses on adequacy. We did not consider fluency because 4 out of the 5 proposed responses to each context were generated by a human. We did not consider informativeness because in the domains considered, it is not necessarily important (in Twitter), or else it seems to correlate highly with adequacy (in Ubuntu).

Context of Conversation A: dearest! question. how many thousands of people can panaad occupy? B: @user panaad has <number> k seat capacity while rizal has <number> k thats why they choose rizal i think .	**Context of Conversation** A: never felt more sad than i am now B: @user aww why ? A: @user @user its a long story ! sure you wanna know it ? bahaha and thanks for caring btw <heart>
Ground Truth Response A: now i know about the siting capacity . thanks for the info @user great evening.	**Ground Truth Response** A: @user i don 't mind to hear it i 've got all day and youre welcome <number>
Proposed Response A: @user makes sense. thanks!	**Proposed Response** A: @user i know , i 'm just so happy for you !

Figure 2: Examples where the metrics rated the response poorly and humans rated it highly (left), and the converse (right). Both responses are given near-zero score by BLEU-N for N> 1. While no metric will perform perfectly on all examples, we present these examples to provide intuition on how example-level errors become aggregated into poor correlation to human judgements at the corpus-level.

responses with BLEU, we recommend the choice of $N = 2$ over $N = 3$ or 4. Note that using a test corpus larger than the size reported in this paper may lead to stronger correlations for BLEU-3 and BLEU-4, due to a higher number of non-zero scores.

It is interesting to note that, while some of the embedding metrics and BLEU show small positive correlation in the non-technical Twitter domain, there is no metric that significantly correlates with humans on the Ubuntu Dialogue Corpus. This is likely because the correct Ubuntu responses contain specific technical words that are less likely to be produced by our models. Further, it is possible that responses in the Ubuntu Dialogue Corpus have intrinsically higher variability (or entropy) than Twitter when conditioned on the context, making the evaluation problem significantly more difficult.

Figure 1 illustrates the relationship between metrics and human judgements. We include only the best performing metric using word-overlaps, i.e. the BLEU-2 score (left), and the best performing metric using word embeddings, i.e. the vector average (center). These plots show how weak the correlation is: in both cases, they appear to be random noise. It seems as though the BLEU score obtains a positive correlation because of the large number of responses that are given a score of 0 (bottom left corner of the first plot). This is in stark contrast to the inter-rater agreement, which is plotted between two randomly sampled halves of the raters (right-most plots). We also calculated the BLEU scores after removing stopwords and punctuation from the responses. As shown in Table 4, this weakens the correlation with human judgements for BLEU-2 compared to the values in Table 3, and suggests that BLEU is sensitive to factors that do not change the semantics of the response.

Finally, we examined the effect of response length on the metrics, by considering changes in scores when the ground truth and proposed response had a large difference in word counts. Table 4 shows that BLEU and METEOR are particularly sensitive to this aspect, compared to the Embedding Average metric and human judgement.

Qualitative Analysis. In order to determine specifically why the metrics fail, we examine qualitative samples where there is a disagreement between the metrics and human rating. Although these only show inconsistencies at the example-level, they provide some intuition as to why the metrics don't correlate with human judgements at the corpus-level. We present in Figure 2 two examples where all of the embedding-based metrics and BLEU-1 score the proposed response significantly differently than the humans.

The left of Figure 2 shows an example where the embedding-based metrics score the proposed response lowly, while humans rate it highly. It is clear from the context that the proposed response is reasonable – indeed both responses intend to express gratitude. However, the proposed response has a different wording than the ground truth response, and therefore the metrics are unable to separate the salient words from the rest. This suggests that the embedding-based metrics would ben-

efit from a weighting of word saliency.

The right of the figure shows the reverse scenario: the embedding-based metrics score the proposed response highly, while humans do not. This is most likely due to the frequently occurring 'i' token, and the fact that 'happy' and 'welcome' may be close together in the embedding space. However, from a human perspective there is a significant semantic difference between the responses as they pertain to the context. Metrics that take into account the context may be required in order to differentiate these responses. Note that in both responses in Figure 2, there are no overlapping n-grams greater than unigrams between the ground truth and proposed responses; thus, all of BLEU-2,3,4 would assign a score near 0 to the response.

6 Discussion

We have shown that many metrics commonly used in the literature for evaluating unsupervised dialogue systems do not correlate strongly with human judgement. Here we elaborate on important issues arising from our analysis.

Constrained tasks. Our analysis focuses on relatively unconstrained domains. Other work, which separates the dialogue system into a dialogue planner and a natural language generation component for applications in constrained domains, may find stronger correlations with the BLEU metric. For example, Wen et al. (2015) propose a model to map from dialogue acts to natural language sentences and use BLEU to evaluate the quality of the generated sentences. Since the mapping from dialogue acts to natural language sentences has lower diversity and is more similar to the machine translation task, it seems likely that BLEU will correlate better with human judgements. However, an empirical investigation is still necessary to justify this.

Incorporating multiple responses. Our correlation results assume that only one ground truth response is available given each context. Indeed, this is the common setting in most of the recent literature on training end-to-end conversation models. There has been some work on using a larger set of automatically retrieved plausible responses when evaluating with BLEU (Galley et al., 2015b). However, there is no standard method for doing this in the literature. Future work should examine how retrieving additional responses affects the correlation with word-overlap metrics.

Searching for suitable metrics. While we provide evidence against existing metrics, we do not yet provide good alternatives for unsupervised evaluation. Despite the poor performance of the word embedding-based metrics in this survey, we believe that metrics based on distributed sentence representations hold the most promise for the future. This is because word-overlap metrics will simply require too many ground-truth responses to find a significant match for a reasonable response, due to the high diversity of dialogue responses. As a simple example, the skip-thought vectors of Kiros et al. (2015) could be considered. Since the embedding-based metrics in this paper only consist of basic averages of vectors obtained through distributional semantics, they are insufficiently complex for modeling sentence-level compositionality in dialogue. Instead, these metrics can be interpreted as calculating the *topicality* of a proposed response (i.e. how on-topic the proposed response is, compared to the ground-truth).

All of the metrics considered in this paper directly compare a proposed response to the ground-truth, without considering the context of the conversation. However, metrics that take into account the context could also be considered. Such metrics could come in the form of an *evaluation model* that is learned from data. This model could be either a discriminative model that attempts to distinguish between model and human responses, or a model that uses data collected from the human survey in order to provide human-like scores to proposed responses. Finally, we must consider the hypothesis that learning such models from data is no easier than solving the problem of dialogue response generation. If this hypothesis is true, we must concede and always use human evaluations together with metrics that only roughly approximate human judgements.

References

R. Artstein, S. Gandhe, J. Gerten, A. Leuski, and D. Traum. 2009. Semi-formal evaluation of conversational characters. In *Languages: From Formal to Natural*, pages 22–35. Springer.

S. Banerjee and A. 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*.

O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58. Association for Computational Linguistics Baltimore, MD, USA.

A. Cahill. 2009. Correlating human and automatic evaluation of a german surface realiser. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 97–100. Association for Computational Linguistics.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *EACL*, volume 6, pages 249–256.

C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53. Association for Computational Linguistics.

C. Callison-Burch, P. Koehn, C. Monz, and O. F. Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Association for Computational Linguistics.

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

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

D. Espinosa, R. Rajkumar, M. White, and S. Berleant. 2010. Further meta-evaluation of broad-coverage surface realization. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 564–574. Association for Computational Linguistics.

P. W. Foltz, W. Kintsch, and T. K. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

G. Forgues, J. Pineau, J.-M. Larcheveque, and R. Tremblay. 2014. Bootstrapping dialog systems with word embeddings.

M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. l, J. Gao, and B. Dolan. 2015a. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Short Papers)*.

M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. Mitchell, J. Gao, and B. Dolan. 2015b. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. *arXiv preprint arXiv:1506.06863*.

Y. Graham, N. Mathur, and T. Baldwin. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proc. of NAACL-HLT*, pages 1183–1191. Citeseer.

A. Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

E. Hovy. 1999. Toward finely differentiated evaluation metrics for machine translation. In *Proceedings of the Eagles Workshop on Standards and Evaluation*.

K. Jokinen and M. McTear. 2009. *Spoken Dialogue Systems*. Morgan Claypool.

C. Kamm. 1995. User interfaces for voice applications. *Proceedings of the National Academy of Sciences*, 92(22):10031–10037.

R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3276–3284.

T. K. Landauer and S. T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

N. Lasguido, S. Sakti, G. Neubig, T. Tomoki, and S. Nakamura. 2014. Utilizing human-to-human conversation examples for a multi domain chat-oriented dialog system. *IEICE TRANSACTIONS on Information and Systems*, 97(6):1497–1505.

J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

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

R. Lowe, N. Pow, I. V. Serban, and J. Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

J. Mitchell and M. Lapata. 2008. Vector-based models of semantic composition. In *ACL*, pages 236–244.

S. Möller, R. Englert, K. Engelbrecht, V. Hafner, A. Jameson, A. Oulasvirta, A. Raake, and N. Reithinger. 2006. MeMo: towards automatic usability evaluation of spoken dialogue services by user error simulations. In *INTERSPEECH*.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002a. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on Association for Computational Linguistics (ACL)*.

K. Papineni, S. Roukos, T. Ward, J. Henderson, and F. Reeder. 2002b. Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results. In *Proceedings of the second international conference on Human Language Technology Research*, pages 132–137.

E. Reiter and A. Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

A. Ritter, C. Cherry, and B. Dolan. 2010. Unsupervised modeling of twitter conversations. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

A. Ritter, C. Cherry, and W. B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics.

V. Rus and M. Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Schatzmann, K. Georgila, and S. Young. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. 2015. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Networks. In *AAAI Conference on Artificial Intelligence*.

I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. 2016. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*.

A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao, and B. Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2015)*.

A. Stent, M. Marge, and M. Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 341–351. Springer.

O. Vinyals and Q. Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

M. Walker, D. Litman, C. Kamm, and A. Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. ACL.

T.-H. Wen, M. Gasic, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

J. Wieting, M. Bansal, K. Gimpel, and K. Livescu. 2015. Towards universal paraphrastic sentence embeddings. *CoRR*, abs/1511.08198.