Relevance in Dialogue: An empirical comparison of existing metrics, and a novel simple metric

Anonymous ACL submission

Abstract

In this work, we evaluate various existing dialogue relevance metrics, find strong dependencies on the dataset, often with poor correlation with human scores of relevance, and propose modifications to reduce data requirements while improving correlation. With these changes, our metric achieves a new state-of-the-art on the HUMOD dataset (Merdivan et al., 2020). We achieve this without fine-tuning, using only 3750 unannotated human dialogues and a single negative example. Despite these limitations, we demonstrate competitive performance on three datasets from different domains. Our code including our metric and data processing is open sourced [1].

1 Introduction

The automatic evaluation of generative dialogue systems remains an important open problem, with potential applications from tourism (Şimşek and Fensel, 2018) to medicine (Fazzinga et al., 2021). In recent years, there has been increased focus on interpretable approaches (Deriu et al., 2021; Chen et al., 2021) often through combining various sub-metrics, each for a specific aspect of dialogue (Berlot-Attwell and Rudzicz, 2021; Phy et al., 2020; Mehri and Eskenazi, 2020). One of these key aspects is "relevance" or "dialogue coherence", commonly defined as whether "[r]esponses are on-topic with the immediate dialogue history" (Finch and Choi, 2020).

These interpretable approaches have motivated measures of dialogue relevance that are not reliant on expensive human annotations. Such measures have appeared in many recent papers on dialogue evaluation, including USR (Mehri and Eskenazi, 2020), USL-H (Phy et al., 2020), and others (Pang et al., 2020; Merdivan et al., 2020). Additionally, dialogue relevance has been used directly in training dialogue models (Xu et al., 2018).

Despite this work, comparison between these approaches has been limited. Aggravating this problem is that authors often collect human annotations on their own datasets with varying amounts and types of non-human responses and, as a result, comparing between approaches has been difficult, if not impossible. We address this problem by evaluating and comparing three prior approaches on three publicly available datasets of dialogue annotated with human ratings of relevance. We find poor correlation with human ratings in various methods, with high sensitivity to dataset.

Based on our observations, we propose a simple metric of logistic regression trained on BERT features (Devlin et al., 2019), using "I don't know." as the only negative example. With this metric, described below, we achieve state-of-the-art correlation on the HUMOD dataset (Merdivan et al., 2020). We make our code, data processing, and empirical setup publicly available to encourage more comparable results in future research.

The primary contributions of this paper are: (i) empiric evidence that current dialogue relevance metrics for English are sensitive to dataset, and often have poor correlation with human ratings, (ii) a simple relevance metric that exhibits good correlation, and (iii) the counter-intuitive result that a single negative example can be equally effective as random negative sampling.

2 Prior metrics

Prior metrics of relevance in dialogue can generally be divided into more traditional approaches that are token-based, and more current approaches based on large pretrained models. These metrics are given the *context* (i.e., the two-person conversation up to a given point in time), as well as a *response* (i.e., the next speaker's response, also known as the 'next turn' in the conversation). From these, they

[1] See Supplemental Material. A Github repository will be made available upon publication.

produce a measure of the response's relevance to the context. Typically, the ground-truth response (also known as the 'gold response') is not assumed to be available.

2.1 n-gram approaches

There have been attempts to use metrics based on n-grams from machine-translation and summarization, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) in dialogue. A significant disadvantage of these approaches is that they rely on access to the ground-truth response, that may not be available (e.g., if the model is being evaluated with self-play). Furthermore, it has been long established that these approaches do not work for measuring dialogue quality (Liu et al., 2016) – this is widely hypothesized to be because a single context can have a wide variety of valid responses. Recent work applied these same methods to dialogue relevance, and found that the correlation with human scores was not significantly different than zero (Merdivan et al., 2020).

2.2 Average-Embedding cosine similarity

Xu et al. (2018) proposed to measure the cosine similarity of a vector representation of the context, and the response. Specifically, the context and response are represented via an aggregate (typically an average) of the uncontextualized word embeddings. This approach can be modified to exploit language models by instead using contextualized word embeddings.

2.3 Fine-tuned embedding model for Next Utterance Prediction (NUP)

This family of approaches combines a word embedding model (typically max- or average-pooled BERT word embeddings) with a simple 1-3 layer MLP, trained for next utterance prediction (typically using negative sampling) (Mehri and Eskenazi, 2020; Phy et al., 2020). The embedding model is then fine-tuned to the domain of interest. In some variants, the model is provided with information in addition to the context and response; e.g., Mehri and Eskenazi (2020) measured relevance on annotated Topical-Chat data (Gopalakrishnan et al., 2019) by appending the topic string to the context. This general architecture and training paradigm have also been directly used as a metric of overall dialogue quality (Ghazarian et al., 2019). In this paper, we focus on the specific implementation by Phy et al. (2020). They use max-pooled BERT embeddings that are passed into a single-layer MLP followed by softmax with two classes. Binary cross-entropy loss and random sampling of negative examples is used at train time.

Note that, for methods that are fine-tuned or otherwise require training, it will often be the case that annotated relevance data is not available on the domain of interest. As a result, the model performance (i.e., correlation with human annotations) cannot be measured on a validation set, and some other means must be used to determine when training must stop (e.g., loss on the surrogate task, or halting after a certain number of epochs). It is therefore important that either the surrogate loss correlates well with the model performance, or the true validation curves of these methods be relatively smooth and monotone so as to reduce the risk of halting training on a model with poor performance.

Another concern with using trained metrics to measure trained dialogue systems is that they may both learn the same patterns in the training data. An extreme example would be a dialogue model that learns only to reproduce responses from the training data verbatim, and a relevance metric that learns to only accept verbatim responses from the training data. We believe that this risk can be reduced by training the metric on separate data from the model (possibly from a different domain). However, unless new training examples can be collected easily, then this approach is only practical if the metric can be trained with a relatively small amount of data and therefore does not compete with the dialogue model for training examples.

2.4 Normalized conditional probability

Pang et al. (2020) also exploited pretrained models, however they instead relied on a generative language model (specifically GPT-2). Their proposed metric is the conditional log-probability of the response given the context, normalized to the range $[0, 1]$. Specifically, for a context q with candidate response r, their proposed relevance score is defined as:

$$c(q \mid r) = -\frac{\max(c_{5th}, \frac{1}{|r|} \log P(r \mid q)) - c_{5th}}{c_{5th}}$$

, where $|r|$ is the number of tokens in the response, $P(r \mid q)$ is the conditional probability of the response given the context under the language model,

and c_{5th} is the 5^{th} percentile of the distribution of $\frac{1}{|r|} \log P(r \mid q)$ over the examples being evaluated.

3 Datasets used for analysis

A literature review reveals that many of these methods have never been evaluated on the same datasets. As such, it is unclear both how these approaches compare, and how well (if at all) they generalize to new data. For this reason, we consider three publicly available English datasets of both human and synthetic dialogue with human annotations of relevance.

All datasets are annotated with Likert ratings of relevance from various reviewers; following Merdivan et al. (2020), we average these ratings over all reviewers. Due to variations in data collection procedures, as well as anchoring effects when rating dialogue (Li et al., 2019), individual Likert ratings from different datasets may not be directly comparable. For this reason, we do not merge the datasets and instead keep them separate. This has the additional benefit of allowing us to observe how methods generalize across datasets.

3.1 HUMOD Dataset

The HUMOD dataset (Merdivan et al., 2020) is an annotated subset of the Cornell movie dialogue dataset (Danescu-Niculescu-Mizil and Lee, 2011). The Cornell dataset consists of $220,579$ conversations from 617 films. The HUMOD dataset is a subset of 4750 contexts, each consisting of at least two and at most seven turns. Every context is paired with both the original human response, and a randomly sampled human response. Each response is annotated with crowd-sourced ratings of relevance from 1-5. The authors measured inter-annotator agreement via Cohen's kappa score (Cohen, 1968), and it was found to be 0.86 between the closest ratings, and 0.42 between randomly selected ratings. Following the authors, we split the dataset into a training set consisting of the first 3750 contexts, a validation set of the next 500 contexts, and a testset of the remaining 500 contexts. As it is unclear how the HUMOD dataset was subsampled from the Cornell movie dialogue dataset, we do not use the Cornell movie dialogue dataset as training data for any of our methods.

3.2 USR Topical-Chat Dataset (USR-TC)

The USR-TC dataset is a subset of the Topical-Chat (TC) dialogue dataset (Gopalakrishnan et al., 2019) created by Mehri and Eskenazi (2020). The Topical-Chat dataset consists of approximately $11,000$ conversations between Amazon Mechanical Turk workers, each grounding their conversation in a provided reading set. The USR-TC dataset consists of 60 contexts taken from the TC frequent test set, each consisting of 1-19 turns. Every context is paired with six responses: the original human response, a newly created human response, and four samples taken from a Transformer dialog model (Vaswani et al., 2017). Each sample follows a different decoding strategy, namely: argmax sampling, and nucleus sampling (Holtzman et al., 2020) at the rates $p = 0.3, 0.5, 0.7$, respectively. Each response is annotated with a human 1-3 score of relevance, produced by one of six dialogue researchers. The authors reported an inter-annotator agreement of 0.56 (Spearman's correlation). We divide the dataset evenly into a validation and test set, each containing 30 contexts. We use the TC train set as the training set.

3.3 Pang et al. (2020) Annotated DailyDialogue Dataset (P-DD)

The P-DD dataset (Pang et al., 2020) is a subset of the DailyDialogue (DD) dataset (Li et al., 2017). The DailyDialogue dataset consists of 13,118 conversations scraped from various websites, specifically digital spaces where English language learners could practice English conversation. The P-DD dataset contains 200 contexts, each consisting of a single turn. Each context is paired with a single synthetic response, generated by a 2-layer LSTM (Bahdanau et al., 2015). Responses are sampled using top-K sampling for $k \in \{1, 10, 100\}$; note that k varies by context. Each response is annotated with ten crowdsourced 1-5 ratings of relevance. The authors reported that inter-annotator Spearman's correlation varied between 0.57 and 0.87. Due to the very small size of the dataset (only 200 dialogues in total), and the lack of information on how the contexts were sampled, we choose to use this dataset exclusively for testing.

4 Evaluating Prior Metrics

For each of the aforementioned datasets, we evaluate:

- COS-FT: an average embedding cosine similarity. Specifically, we use the implementation[2] provided by Csáky et al. (2019). This

[2] https://github.com/ricsinaruto/

	HUMOD		USR-TC		P-DD	
Prior Metric	S	P	S	P	S	P
COS-FT	0.09	0.10	*0.26	*0.24	−0.10	−0.11
COS-MAX-BERT	*0.13	*0.10	*0.20	0.14	0.03	0.02
COS-NSP-BERT	0.08	0.06	0.08	0.09	*0.30	*0.23
NORM-PROB	*0.19	*0.16	*−0.24	*−0.26	*0.65	*0.59
NUP-BERT (H)	*0.33 (0.02)	*0.37 (0.02)	0.10 (0.02)	*0.22 (0.01)	*0.62 (0.03)	*0.54 (0.02)
NUP-BERT (TC-S)	*0.29 (0.02)	*0.35 (0.03)	†0.17 (0.03)	†0.20 (0.04)	*0.58 (0.05)	*0.56 (0.04)
NUP-BERT (TC)	*0.30 (0.01)	*0.38 (0.00)	0.16 (0.02)	*0.21 (0.02)	*0.62 (0.04)	*0.58 (0.03)

Table 1: Spearman (S) and Pearson (P) correlations of baseline models with average human ratings on the test sets (correlations on the validation set can be found in the Appendix, Table 6). Models with a trained or fitted component (i.e., NUP-BERT variants) are averaged over three runs, with the standard deviation reported in brackets. They also have their training data specified in brackets, (H) signifies HUMOD, (TC-S) signifies a subset of TC containing 3750 dialogues (same size as the HUMOD train set), and (TC) signifies the full Topical Chat training set. A correlation is marked with '*' if all trials were significant at the $p < 0.01$ level. Otherwise, a correlation is marked with '†' if at least one trial was significant at the $p < 0.01$ level. Note that the COS-FT and NORM-PROB baselines attain negative correlation with human scores on the P-DD and USR-TC datasets respectively.

	HUMOD		USR-TC		P-DD	
Metric	S	P	S	P	S	P
NUP-BERT (H)	*0.33 (0.02)	*0.37 (0.02)	0.10 (0.02)	*0.22 (0.01)	***0.62** (0.03)	*0.54 (0.02)
NUP-BERT (TC-S)	*0.29 (0.02)	*0.35 (0.03)	†0.17 (0.03)	†0.20 (0.04)	*0.58 (0.05)	*0.56 (0.04)
NUP-BERT (TC)	*0.30 (0.01)	*0.38 (0.00)	0.16 (0.02)	*0.21 (0.02)	***0.62** (0.04)	***0.58** (0.03)
IDK (H)	***0.58** (0.00)	***0.58** (0.00)	**0.18** (0.00)	***0.24** (0.00)	*0.53 (0.00)	*0.48 (0.01)
IDK (TC-S)	***0.58** (0.00)	***0.58** (0.00)	**0.18** (0.00)	*0.22 (0.00)	*0.54 (0.01)	*0.49 (0.01)

Table 2: Comparison of our proposed metric against the NUP-BERT baseline on the test set (corresponding correlations on the validation set can be found in Table 7). Note the strong improvement on HUMOD and equivalent, or slightly improved performance on USR-TC, at the cost of performance loss on P-DD.

implementation uses average fastText [3] embeddings.

- COS-MAX-BERT: another cosine similarity. For better comparison with BERT-based approaches, and inspired by BERT-RUBER (Ghazarian et al., 2019), we instead use max-pooled BERT contextualized word embeddings.
- COS-NSP-BERT: another cosine similarity embedding modified to use BERT, specifically altered to use the pretrained features extracted from the [CLS] token for the pretrained next-sentence-prediction head.
- NUP-BERT: a fine-tuned BERT next-utterance prediction approach. Specifically, we use the NUP score implementation[4] provided by Phy et al. (2020). We experiment with fine-tuning BERT to the HUMOD test set (3750 dialogues), the full TC test set, and TC-S (a subset of the TC training set containing only 3750 dialogues).
- NORM-PROB: a GPT-2 based normalized conditional-probability approach. Specifically, we use the implementation[5] provided by Pang et al. (2020). Note that the P-DD dataset was released in the same paper.

In all cases, we use hugging-face `bert-base-uncased` as the pretrained BERT model. Only NUP-BERT was fine-tuned. To prevent an unfair fitting to any specific dialogue model, and to better reflect the evaluation of a new dialogue model, only human responses were used at train time. All hyperparameters were left at their recommended values. Note that test-set performance is averaged over 3 runs for NUP-BERT.

Also note that n-gram approaches were not evaluated. This is in part due to previous evidence suggesting no correlation (Merdivan et al., 2020), and, in part, as these methods require the gold-truth reference for comparison. As a result, these methods cannot be used to fairly evaluate the gold-truth response. Since our annotated datasets are relatively small and contain limited amounts of human

dialog-eval

[3] https://fasttext.cc/

[4] https://github.com/vitouphy/usl_dialogue_metric

[5] https://github.com/alexzhou907/dialogue_evaluation

generated responses, we decided that discarding the annotated ground-truth responses would harm our ability to evaluate the other metrics. Furthermore, the P-DD dataset does not include ground-truth human responses, thereby making evaluation on this dataset impossible.

4.1 Analysis

Table 1 makes it immediately clear that the normalized probability (NORM-PROB) and cosine similarity (COS-FT, COS-MAX-BERT, COS-NSP-BERT) approaches do not generalize well across datasets. Although NORM-PROB works very well on the P-DD dataset, it has weak performance on HUMOD and has, in fact, a significant *negative* correlation on USR-TC. As this metric was developed for the P-DD dataset, and the P-DD dataset consists solely of synthetic responses from an LSTM model, we believe that this approach is over-fitted to measuring relevance on this LSTM model. Similarly, although the cosine-similarity approach using FastText word embeddings has the best performance on the USR-TC dataset, it performs poorly on HUMOD, and has negative correlation on P-DD. As such, it is clear that, while both cosine-similarity and normalized probability approaches can perform well, they have serious limitations. They are very sensitive to the domain and models under evaluation, and are capable of becoming negatively correlated with human ratings under suboptimal conditions.

The final baseline, NUP-BERT, appears to have the best overall performance, outperforming each of the other baselines on at least 2 of the datasets. Despite this, we can see that performance on HUMOD and USR-TC is still fairly weak. We can also observe that although fine-tuning on HUMOD data results in lower Spearman's correlation on USR-TC, the general trend is that performance tends to be comparable regardless of training data. This appears to be true both with respect to domain (H vs TC) and amount of training data (TC vs TC-S).

Overall, the results of Table 1 are concerning as they suggest that at least two current approaches generalize poorly across either dialogue models, or domains. As a result, research into new dialogue relevance metrics is required. Furthermore, it is clear that the methodology for the evaluation of dialogue relevance metrics must be updated to use various dialogue models in various different domains.

5 IDK: A novel metric for dialogue relevance

Based on these results, we propose a number of modifications to the NUP-BERT metric to produce a novel metric that we call IDK ("I don't know"). The overall architecture is mostly unchanged, however the training procedure and the exact features used are altered.

First, based on the observation that the amount of training data has little impact, we decide to freeze BERT features entirely and do not fine-tune to the domain. Instead, we focus on the next-utterance-prediction task. More specifically, whereas the NUP-BERT baseline uses max-pooled BERT word embeddings, we instead use the pre-trained next sentence prediction features – from the huggingface v2.11.0 documentation [6]: "(classification token) further processed by a Linear layer and a Tanh activation function. The Linear layer weights are trained from the next sentence prediction (classification) objective during pre-training".

Second, to improve generalization and reduce variation in training (particularly important as the practitioner likely does not have access to annotated relevance data), and operating on the assumption that relevance is captured by a few key dimensions of the NUP features, we add L1 regularization to our regression weights ($\lambda = 1$).

Third, in place of random sampling we use a fixed negative sample, "i don't know", at all time steps. This allows us to train the model on less data, as a corpus from which to sample negative samples is no longer required.

Additionally, we perform a minor simplification of the model to reduce the number of weights, using logistic regression in place of 2-class softmax. We train for 2 epochs using binary cross-entropy loss – the same as the NUP-BERT baseline. We use the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.001.

Table 2 reports the correlation between the metric's responses and the average human rating. We achieve a Pearson's correlation on HUMOD of 0.58 which, to our knowledge, represents a new state-of-the-art performance for reference-less metrics on HUMOD. This performance surpasses the previous SOTA (0.138 with HAN-R(CE) (Merdivan et al., 2020)), as well as our provided baselines. This performance is also very close to the supervised

[6] https://huggingface.co/transformers/v2.11.0/model_doc/bert.html

SOTA of 0.602 using a supervised fine-tuned BERT model (Merdivan et al., 2020).

We have included some examples of the our metric's output on the HUMOD dataset in Table 3, and we have included a scatter plot of IDK vs human scores in the Appendix, Figures 1-3.

Compared to our baselines, we see that our proposed metric has strong improvement on the HUMOD dataset and equivalent or stronger performance on USR-TC, at a cost of reduced performance on P-DD. As the performance drop on P-DD is less than the performance gain on HUMOD, and as HUMOD is human data rather than LSTM data, we consider this tradeoff to be a net benefit. However, this reinforces our previous conclusion that these approaches are highly sensitive to the dataset. The absolute performance of all metrics studied, including our own, vary considerable by dataset. Furthermore, even the relative performance of closely related metrics such as IDK and NUP-BERT, or COS-FT and COS-NSP-BERT, varies considerably between datasets.

It is worth noting that our approach does not out-perform the cosine and normalized-conditional probability baselines in *all* cases – only the *majority* of cases. As such, when annotated human data is not available for testing, it would appear that our approach is the preferred choice.

5.1 Ablation tests

Table 4 outlines correlation when ablating the L1 regularization, or when using randomly sampled negative examples in place of "i don't know". Specifically, we produce negative examples by shuffling the responses of the next 3750 dialogues in the dataset.

The clearest observation is that L1 regularization is critical to good performance when using "i don't know" in place of negative samples – otherwise, the model presumably overfits. Second, we can see that using "i don't know" in place of negative samples has a mixed, but relatively minor effect. Thirdly we can see that the effect of L1 regularization is quite positive when training on TC data (regardless of the negative samples), and mixed but smaller when training on HUMOD data.

Overall, this suggests that when annotated relevance data is not available, then L1 regularization may be helpful. Its effect varies by domain, but appears to have a much stronger positive effect than a negative effect.

Dialogue Turn	Human	IDK
Mommy –	-	-
Yes, darling.	-	-
Did you ever make a wish?	-	-
Oh, lots of times.	-	-
Did your wishes ever come true?	5.00	4.97
What's your real name?	1.00	3.81
Sometimes.	4.67	4.60
From high school Mary? Yeah, I saw her about six months ago at a convention in Las Vegas.	1.00	1.13
I made a wish today, and it came true just like Edward said it would.	5	4.9
When I am sure I am among friends.	2.33	3.01
Yes, Albert?	-	-
John, we're going huntin'.	-	-
Who's goin?	-	-
We're all going.	-	-
Nick's going?	4.67	4.65
I will keep you safe. We are both older.	2.00	1.09
Nick , Vince , Albert and John.	4.00	4.95
A ride? Hell, that's a good idea. Okay, let's go. Hey, let's go.	2.33	4.68
No women?	4.00	2.39
I guess so	3.00	2.59

Table 3: Two multi-turn examples from HUMOD test set. The randomly sampled distractor turns are italicized, and are not part of the context in subsequent turns. For ease of comparison, IDK scores' range was linearly shifted and re-scaled to 1-5. These scores were generated using IDK trained on HUMOD.

Looking at the table, we can also see that, when combined with L1 regularization, performance using "i don't know" in place of random negative samples is typically slightly lower, but comparable. Therefore, this approach seems to be most appropriate when data is scarce.

5.2 Additional Experiments: Triplet Loss

An intuitive limitation of using "i don't know" as a negative example with binary-cross-entropy loss is that this encourages the model to always map "i don't know" to exactly zero. However, the relevance of "i don't know" evidently varies by context. Clearly, it is a far less relevant response to "I was in-

			HUMOD		USR-TC		P-DD	
Data	L1	IDK	S	P	S	P	S	P
H	✓	✓	*0.58 (0.00)	*0.58 (0.00)	0.18 (0.00)	*0.24 (0.00)	*0.53 (0.00)	*0.48 (0.01)
H		✓	*0.42 (0.06)	*0.42 (0.05)	*0.24 (0.00)	*0.25 (0.00)	*0.29 (0.06)	*0.32 (0.03)
H	✓		*0.61 (0.00)	*0.61 (0.00)	0.12 (0.00)	*0.21 (0.01)	*0.55 (0.00)	*0.52 (0.01)
H			*0.60 (0.00)	*0.61 (0.00)	0.18 (0.00)	*0.26 (0.01)	*0.54 (0.00)	*0.50 (0.01)
TC-S	✓	✓	*0.58 (0.00)	*0.58 (0.00)	0.18 (0.00)	*0.22 (0.00)	*0.54 (0.01)	*0.49 (0.01)
TC-S		✓	*0.36 (0.04)	*0.34 (0.05)	0.17 (0.01)	0.11 (0.01)	*0.34 (0.03)	*0.32 (0.04)
TC-S	✓		*0.59 (0.01)	*0.54 (0.03)	†0.18 (0.04)	*0.27 (0.02)	*0.52 (0.03)	*0.43 (0.05)
TC-S			*0.35 (0.07)	*0.41 (0.01)	†0.13 (0.10)	*0.21 (0.03)	†0.23 (0.10)	†0.27 (0.11)

Table 4: Test correlation of various ablations of the proposed metric. The L1 column signifies whether L1 regularization is used (with $\lambda = 1$), and the IDK column indicates whether the negative samples are "i don't know", or a random shuffle of 3750 other human responses. Note that L1 regularization is beneficial when training on TC-S data (particularly if generalization to other domains is required).

terrupted all week and couldn't get anything done, it was terrible!" than it is to "what is the key to artificial general intelligence?" Motivated by this intuition, we experimented with a modified triplet loss:

$$f_t(c, r) = \max\left(y(c, r) - y(c, r') + m, 0\right)$$

$$\mathcal{L}(c, r) = -\log\left(1 + m - f_t(c, r)\right)$$

Intuitively, a triplet loss would allow for the relevance of "i don't know" to shift, without impacting the loss as long as the ground-truth responses continue to score sufficiently higher. Note that the loss is modified to combat gradient saturation due to the sigmoid non-linearity. However, the results (see Table 5) suggest equivalence, at best. Often, this loss performs equivalently to binary cross-entropy (BCE) but it can also produce degenerate solutions (note the high variance when training on TC data). Furthermore, it does not appear to produce superior correlations.

For this reason, we believe that, although adapting triplet loss for next-utterance prediction in place of binary cross-entropy could be made to work, it does not appear to provide any advantages. If validation data is available, it can be used to confirm whether the model has reached a degenerate solution, and thus this loss could be used interchangeably with BCE. However, there does not appear to be any advantage in doing so.

6 Related Work

In addition to the prior metrics already discussed, the area of dialogue relevance is both motivated by, and jointly developed with, the problem of automatic dialogue evaluation. As relevance is a major component of good dialogue, developments flow from one problem to the other, and vice versa. The NUP-BERT relevance metric is very similar to BERT-RUBER (Ghazarian et al., 2019); both train a small MLP to perform the next-utterance-prediction task based on aggregated BERT features. Both of these share a heritage with earlier attempts to evaluate dialogue using self-supervised methods, such as adversarial approaches to dialogue evaluation that train a classifier to distinguish human from generated samples (Kannan and Vinyals, 2017). Another example of shared development is the use of word-overlap metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) that have been imported wholesale into both dialogue relevance and overall quality from the fields of machine-translation and summarization, respectively.

Simultaneously, metrics of dialogue relevance have been motivated by dialogue evaluation. There is a long history of evaluating dialogue models on various aspects of the overall quality; Finch and Choi (2020) performed a meta-analysis of prior work, and proposed dimensions of: grammaticality, relevance, informativeness, emotional understanding, engagingness, consistency, proactivity, and satisfaction. New approaches to dialogue evaluation have emerged from this body of work, seeking to aggregate individual measures of various dimensions of dialogue, often including relevance (Mehri and Eskenazi, 2020; Phy et al., 2020; Berlot-Attwell and Rudzicz, 2021).

Another connection between these tasks is that they share common problems such as the diversity of valid responses. Furthermore, our findings that existing relevance metrics generalize poorly to new domains is consistent with previous findings in dialogue evaluation. Prior trained automatic dialogue evaluation metrics such as ADEM (Lowe et al., 2017) have been found to generalize poorly

			HUMOD		USR-TC		P-DD	
Data	L1	IDK	S	P	S	P	S	P
H	✓	✓	*0.59 (0.01)	*0.55 (0.02)	0.17 (0.01)	*0.28 (0.01)	*0.54 (0.03)	*0.44 (0.02)
H		✓	*0.15 (0.05)	*0.19 (0.06)	†0.19 (0.01)	*0.25 (0.02)	0.10 (0.04)	†0.17 (0.05)
H	✓		*0.45 (0.24)	*0.42 (0.21)	0.14 (0.04)	†0.23 (0.10)	†0.39 (0.21)	*0.34 (0.14)
H			*0.61 (0.00)	*0.60 (0.01)	0.17 (0.00)	*0.23 (0.01)	*0.55 (0.01)	*0.53 (0.01)
TC-S	✓	✓	*0.32 (0.44)	*0.25 (0.55)	0.12 (0.06)	†0.10 (0.24)	*0.24 (0.47)	*0.21 (0.46)
TC-S		✓	*0.27 (0.11)	*0.26 (0.10)	0.16 (0.02)	0.14 (0.03)	†0.22 (0.12)	†0.22 (0.09)
TC-S	✓		*-0.20 (0.69)	*-0.20 (0.65)	-0.03 (0.17)	†-0.05 (0.29)	*-0.18 (0.62)	*-0.19 (0.54)
TC-S			†0.18 (0.20)	*0.18 (0.06)	0.04 (0.07)	0.09 (0.17)	0.10 (0.07)	0.07 (0.06)

Table 5: Repeat of ablation experiments, however using modified triplet loss ($m = 0.4$) in place of binary cross entropy. Contrary to our intuition, we do not find any improvement in performance. Comparing against Table 4, we find either equivalent or degraded performance, with an additional tendency to converge to a degenerate solution (e.g., see high variances in TC-S with L1 and IDK).

to other domains, or even datasets (Lowe, 2019). Our work suggests that this challenge extends to the subproblem of dialogue relevance as well.

7 Discussion

Our experiments demonstrate that several published measures of dialogue relevance have poor, or even negative, correlation when evaluated on new datasets of dialogue relevance, suggesting overfitting to either model or domain. As such, it is clear that further research into new measures of dialogue relevance is required, and that great care must be taken in their evaluation to compare against a number of different models in a number of domains. Furthermore, it is also clear that for the current practitioner who requires a measure of relevance, there are no guarantees that current methods will perform well on a given domain. As such, it appears to be wise to collect a small validation dataset of human-annotated relevance data for use in selecting a relevance metric. If good correlation is not imperative, then a NUP-based approach such as our outlined metric appears to be the best option for achieving acceptable correlation, even if not trained on the same domain.

When training data is scarce, our results suggest that the use of strong regularization allows for the use of a single negative example, "i don't know", in the place of randomly sampled negative samples. Additionally, our results appear to suggest that the performance of metrics based on NUP-BERT is fairly agnostic to the domain of the training data. As such, training data can be used from a different dialogue domain in place of the domain of interest.

Having said this, it is clear that further research into what exactly these metrics are measuring, and why they fail to generalize, is clearly merited. As an example, although our empiric results suggest that use of a single negative example generalizes across domains, there is no compelling theoretical reason why this should be the case. More generally, all the metrics outlined are complex, dependent on large corpora of text, and, due to cost, created without access to the target task of relevance. As a result, they are all dependent on either surrogate tasks (i.e., NUP), or unsupervised learning (e.g., FastText embeddings). Consequently, it is especially difficult to conclude what exactly these metrics are measuring. At present, the only strong justification that these metrics are indeed measuring relevance is good correlation with human judgements – and poor generalization across similar domains is not an encouraging result.

Although the metric outlined is not appropriate for final model evaluation (as it risks unfairly favouring dialogue models based on the same pretrained BERT, or similar architectures), our aim is that it will prove useful for rapid prototyping and hyperparameter search. Additionally, it is our hope that our findings on the domain sensitivity of existing metrics will spur further research into both the cause of – and solutions to – this problem.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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*, pages 65–72, Ann Ar-

bor, Michigan. Association for Computational Linguistics.

Ian Berlot-Attwell and Frank Rudzicz. 2021. On the use of linguistic features for the evaluation of generative dialogue systems. *CoRR*, abs/2104.06335.

Zhang Chen, João Sedoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. 2021. DSTC10: Track 5: Automatic evaluation and moderation of open-domain dialogue systems. Accessed: 9-7-2021 https://drive.google.com/file/d/1B2YBtWaLJU5X3uudSZEaOyNWQ_QoTZLG/view.

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

Richárd Csáky, Patrik Purgai, and Gábor Recski. 2019. Improving neural conversational models with entropy-based data filtering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5650–5669, Florence, Italy. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bettina Fazzinga, Andrea Galassi, and Paolo Torroni. 2021. An argumentative dialogue system for covid-19 vaccine information.

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Anjuli Kannan and Oriol Vinyals. 2017. Adversarial evaluation of dialogue models. *CoRR*, abs/1701.08198.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Ryan Lowe. 2019. A retrospective for "Towards an automatic Turing test - learning to evaluate dialogue responses". *ML Retrospectives*.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Erinc Merdivan, Deepika Singh, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist. 2020. Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3).

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3981–3991, Brussels, Belgium. Association for Computational Linguistics.

Umutcan Şimşek and Dieter Fensel. 2018. Now we are talking! Flexible and open goal-oriented dialogue systems for accessing touristic services. *e-Review of Tourism Research*.

A Scatter Plots

Figures 1, 2, and 3 illustrate IDK vs human scores of relevance, where the IDK training data is HUMOD. A regression line is fitted to highlight the trend.

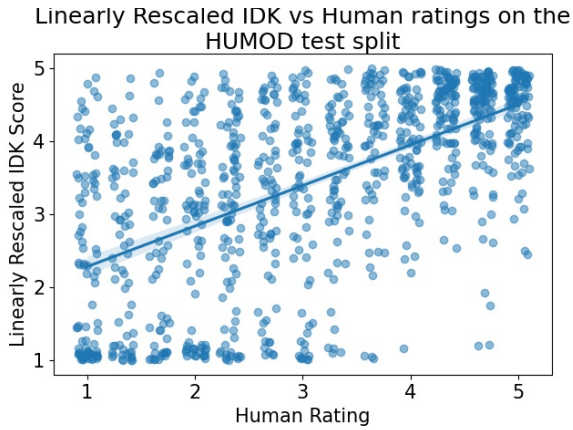

Figure 1: IDK scores, linearly re-scaled to the range 1-5, versus human scores of relevance, on the HUMOD test set.

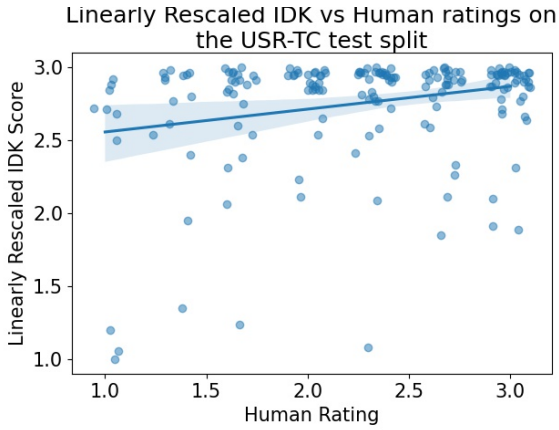

Figure 2: IDK scores, linearly re-scaled to the range 1-3, versus human scores of relevance, on the USR-TC test set.

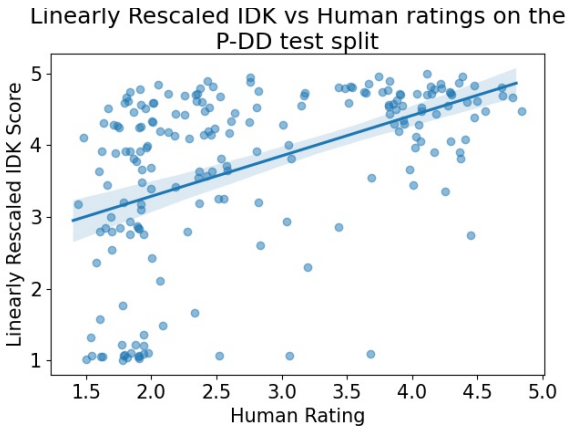

Figure 3: IDK scores, linearly re-scaled to the range 1-5, versus human scores of relevance, on the P-DD test set.

B Performance on validation data split

Correlations of the models on the validation set are outlined in Table 6 for prior metrics, and in Table 7 for all ablations and variants of our model.

	HUMOD		USR-TC	
Prior Metric	S	P	S	P
COS-FT	0.08	0.08	*0.27	0.17
COS-MAX-BERT	0.08	0.05	0.18	*0.19
COS-NSP-BERT	0.06	*0.09	*0.23	*0.25
NORM-PROB	*0.27	*0.25	*−0.29	*−0.30
NUP-BERT (H)	*0.37 (0.01)	*0.38 (0.00)	*0.38 (0.02)	*0.39 (0.01)
NUP-BERT (TC-S)	*0.32 (0.01)	*0.36 (0.02)	*0.38 (0.04)	*0.41 (0.04)
NUP-BERT (TC)	*0.33 (0.02)	*0.37 (0.02)	*0.45 (0.07)	*0.44 (0.02)

Table 6: Spearman (S) and Pearson (P) correlations of prior metrics with human ratings on the validation splits of all provided dataset. As NUP-BERT is trained we perform 3 runs, reporing the mean and standard deviation. (*) denotes $p < 0.01$ accross all trials. Underline indicates a negative correlation. The inter-rater correlation (average correlation of one rater versus the average of all other raters) is also reported. NOTE: USR scores are human only for all excet COS-BERT's

Name	HUMOD Spear	HUMOD Pear	TC Spear	TC Pear
H_IDK_L1	*0.56 (0.01)	*0.53 (0.02)	*0.45 (0.03)	*0.44 (0.02)
H_IDK_bce_L1	*0.57 (0.00)	*0.56 (0.00)	*0.42 (0.01)	*0.41 (0.00)
H_IDK_bce	*0.39 (0.05)	*0.40 (0.05)	*0.36 (0.02)	*0.34 (0.00)
H_IDK	*0.15 (0.05)	*0.19 (0.06)	0.09 (0.05)	†0.21 (0.05)
H_Rand3750_L1	*0.42 (0.22)	*0.40 (0.20)	*0.44 (0.00)	*0.45 (0.01)
H_Rand3750_bce_L1	*0.58 (0.00)	*0.58 (0.00)	*0.45 (0.00)	*0.46 (0.00)
H_Rand3750_bce	*0.58 (0.00)	*0.57 (0.01)	*0.46 (0.00)	*0.43 (0.02)
H_Rand3750	*0.58 (0.00)	*0.58 (0.00)	*0.46 (0.00)	*0.45 (0.02)
TC-S_IDK_L1	*0.29 (0.43)	*0.23 (0.53)	*0.39 (0.07)	*0.41 (0.07)
TC-S_IDK_bce_L1	*0.57 (0.00)	*0.56 (0.00)	*0.43 (0.00)	*0.40 (0.00)
TC-S_IDK_bce	*0.35 (0.04)	*0.33 (0.05)	*0.40 (0.01)	*0.31 (0.01)
TC-S_IDK	*0.25 (0.10)	*0.24 (0.10)	*0.34 (0.05)	*0.36 (0.03)
TC-S_Rand3750_L1	*-0.19 (0.67)	*-0.20 (0.63)	*-0.13 (0.52)	*-0.14 (0.50)
TC-S_Rand3750_bce_L1	*0.56 (0.01)	*0.52 (0.03)	*0.44 (0.03)	*0.40 (0.02)
TC-S_Rand3750_bce	*0.31 (0.05)	*0.36 (0.03)	†0.16 (0.29)	†0.18 (0.26)
TC-S_Rand3750	†0.15 (0.17)	*0.11 (0.02)	†-0.14 (0.24)	†-0.06 (0.27)

Table 7: Validation correlation of all of tested variants and ablations of our model. H vs. TC-S indicates training set (HUMOD or subset of TopicalChat respectively). IDK vs. Rand3750 indicates whether negative examples where "i don't know" or random. If bce is present, then binary cross entropy was used as the loss, otherwise our modified triplet loss is used. If L1 is present, then L1 regulaization with $\lambda = 1$ is used, otherwise no regularization is used. Again, standard deviation over three trials is reported in parentheses, and '*' is used to indicate that all trials were significant at $p < 0.01$. '†' indicates at least on trial was significantly different from zero at $p < 0.01$