Relevance in Dialogue: Is Less More? 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 dependency on the dataset, often with poor correlation with human scores of relevance, and propose modifications to reduce data requirements and domain sensitivity while improving correlation. Our proposed metric achieves state-of-the-art performance on the HUMOD dataset (Merdivan et al., 2020) while reducing measured sensitivity to dataset by 37%-66%. 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 four datasets from different domains. Our code, including our metric and experiments, 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, 2020b). One of these key aspects is "relevance" (sometimes called "context 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, 2020b), 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. Consequently, direct comparisons are not possible. It is known that metrics of dialogue *quality* often perform poorly on new test sets of quality ratings (Yeh et al., 2021), but it remains an open question whether poor generalization also plagues the much simpler dialogue relevance task. We address this problem by evaluating and comparing six prior approaches on four publicly available datasets of dialogue annotated with human ratings of relevance. We find poor correlation with human ratings across various methods, with high sensitivity to dataset.

Based on our observations, we propose a simple metric of logistic regression trained on pretrained BERT NSP features (Devlin et al., 2019), using "i don't know." as the only negative example. With this metric, we achieve state-of-the-art correlation on the HUMOD dataset (Merdivan et al., 2020). We release our metric and evaluation code to encourage comparable results in future research.

Our primary contributions are: (i) empirical 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 reduced domain sensitivity, 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

[1] See Supplemental; code will be made available on GitHub

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 produce a measure of the response's relevance to the context. The ground-truth response (i.e., the 'gold response') may or may not 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. However, we discard these approaches due to their limitations: they require a ground-truth response, and correlate poorly with dialogue relevance (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, 2020b; 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 (2020b) appended a topic string to the context. This approach has 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): max-pooled BERT embeddings passed into a single-layer MLP followed by two-class softmax, trained with binary cross-entropy (BCE) loss and random sampling of negative samples.

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, model performance cannot be measured on a validation set during training. A different method must be employed 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. However, 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. Alternatively, a sufficiently generalizable metric may be trained on data from a different domain.

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 $[0, 1]$ (see Appendix D.1 for details).

Mehri and Eskenazi (2020a) also relied on a generative language model (specifically, DialoGPT (Zhang et al., 2020)), however their approach measured the probability of followup-utterances, e.g., "Why are you changing the topic?" to indicate irrelevance. Their relevance and correctness scores are defined as $c(q \mid r) = -\sum_{i=1}^{|n|} \log P(n_i \mid r, q)$, where $n_i \in n$ is a negative response suggesting irrelevance or incorrectness. Note that positive utterances can be used, however the author's measures of correctness and relevance only used negative utterances.

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 they generalize to new data. For this reason, we consider four publicly available English datasets of both human and synthetic dialogue with human relevance annotations. All datasets are annotated with Likert ratings of relevance from multiple 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 (Li et al., 2019), Likert ratings from different datasets may not be directly comparable. Consequently, we keep the datasets separate. This also allows us to observe generalization across datasets.

Altogether, our selected datasets cover a wide variety of responses, including human, LSTM, Transformer, Meena (Adiwardana et al., 2020), and Mitsuku[2] generated responses, and random distractors.

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 between two and 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 test-set of the remaining 500 contexts. As it is unclear how HUMOD was subsampled from the Cornell movie dialogue dataset, we do not use the Cornell movie dialogue dataset as training data.

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 (2020b). 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 websites where English language learners could practice English conversation. The P-DD dataset contains 200 contexts, each of a single turn and 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 with a reported inter-annotator Spearman's correlation 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 use this dataset exclusively for testing.

3.4 FED Dataset

The FED dataset (Mehri and Eskenazi, 2020a), consists of 375 annotated dialogue turns taken from 40 human-human, 40 human-Meena (Adiwardana et al., 2020), and 40 human-Mitsuku conversations. We use a subset of the annotations, specifically turnwise relevance, and turnwise correctness (the latter defined by the authors as whether there was a "a misunderstanding of the conversation"). As the authors note, their definition of correctness is often encapsulated within relevance; we thus evaluate on both annotations. Due to the small size, we used this dataset only for testing.

4 Evaluating Prior Metrics

For each of the aforementioned datasets, we evaluate the following relevance metrics:

- COS-FT: average fastText[3] embedding cosine similarity. Code by Csáky et al. (2019)
- COS-MAX-BERT: Cosine similarity with max-pooled BERT contextualized word em-

[2] 2019 Loebner prize winning system

[3] https://fasttext.cc/

	HUMOD		USR-TC		P-DD		FED-Correctness		FED-Relevance	
Prior Metric	S	P	S	P	S	P	S	P	S	P
COS-FT	0.09	0.10	*0.26	*0.24	−0.02	−0.04	0.08	0.04	0.11	0.07
COS-MAX-BERT	*0.13	*0.10	*0.20	0.14	0.03	0.02	0.03	0.01	0.06	0.04
COS-NSP-BERT	0.08	0.06	0.08	0.09	*0.30	*0.23	−0.03	−0.01	−0.04	−0.02
NORM-PROB	*0.19	*0.16	*−0.24	*−0.26	*0.65	*0.59	0.05	0.06	0.07	0.07
FED-CORRECT	−0.06	−0.04	−0.08	−0.12	*−0.25	*−0.26	*0.17	*0.17	*0.15	*0.15
FED-RELEVANT	−0.06	−0.05	−0.08	−0.12	*−0.26	*−0.27	*0.17	*0.17	*0.15	*0.15
GRADE	*0.61	*0.61	0.00	0.03	*0.70	*0.68	0.12	0.12	*0.15	*0.15
DYNA-EVAL	*0.09	*0.10	0.10	0.10	0.00	-0.02	*0.26	*0.27	*0.32	*0.31
NUP-BERT (H)	*0.33 (0.02)	*0.37 (0.02)	0.10 (0.02)	*0.22 (0.01)	*0.62 (0.04)	*0.54 (0.02)	†0.14 (0.04)	*0.21 (0.03)	*0.22 (0.01)	*0.30 (0.01)
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)	0.05 (0.04)	0.12 (0.01)	†0.16 (0.04)	*0.21 (0.01)
NUP-BERT (TC)	*0.30 (0.01)	*0.38 (0.00)	0.16 (0.02)	*0.21 (0.02)	*0.62 (0.05)	*0.58 (0.04)	0.06 (0.01)	†0.12 (0.02)	*0.18 (0.02)	*0.23 (0.01)

Table 1: Spearman (S) and Pearson (P) correlations of baseline models with average human ratings on the test sets. BERT-NUP is averaged over three runs, with the standard deviation reported in brackets. Training data is specified in brackets: (H) signifies HUMOD, (TC) signifies the Topical Chat training set, and (TC-S) signifies a subset of TC containing 3750 dialogues (same size as the HUMOD train set). '*' indicates all trials were significant at the $p < 0.01$ level. '†' indicates at least one trial was significant. Note that most cosine and language-model based metrics attain negative correlation with human scores.

	HUMOD		USR-TC		P-DD		FED-Correctness		FED-Relevance	
Prior Metric	S	P	S	P	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** (0.04)	*0.54 (0.02)	†0.14 (0.04)	*0.21 (0.03)	*0.22 (0.01)	***0.30** (0.01)
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)	0.05 (0.04)	0.12 (0.01)	†0.16 (0.04)	*0.21 (0.01)
NUP-BERT (TC)	*0.30 (0.01)	*0.38 (0.00)	0.16 (0.02)	*0.21 (0.02)	***0.62** (0.05)	***0.58** (0.04)	0.06 (0.01)	†0.12 (0.02)	*0.18 (0.02)	*0.23 (0.01)
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)	***0.15** (0.00)	***0.23** (0.00)	***0.24** (0.00)	*0.29 (0.00)
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)	***0.15** (0.00)	***0.23** (0.00)	***0.24** (0.00)	*0.29 (0.00)

Table 2: Comparison of our proposed metric (IDK) against the NUP-BERT baseline on the test set. Note the strong improvement on HUMOD and equivalent, or slightly improved performance on USR-TC, at the cost of performance loss on P-DD. Note IDK performance is almost independent of training data.

beddings, inspired by BERT-RUBER (Ghazarian et al., 2019)

- COS-NSP-BERT: Cosine similarity using the pretrained features extracted from the [CLS] token used by next-sentence-prediction head.
- NUP-BERT: Fine-tuned BERT next-utterance prediction approach. Implementation by Phy et al. (2020). We experiment with fine-tuning BERT to the HUMOD train set (3750 dialogues), the full TC train set, and TC-S (a subset of the TC training set containing 3750 dialogues).
- NORM-PROB: GPT-2 based normalized conditional-probability; approach and implementation by Pang et al. (2020); note that the P-DD dataset was released in the same paper.
- FED-RELEVANT & FED-CORRECT: DialoGPT based normalized conditional-probability; approach and implementation by Mehri and Eskenazi (2020a)

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. NUP-BERT performance is averaged over 3 runs.

Note that we also evaluate GRADE (Huang et al., 2020) and DYNA-EVAL (Zhang et al., 2021); however these do not measure relevance, but rather *dialogue coherence*: "whether a piece of text is in a consistent and logical manner, as opposed to a random collection of sentences" (Zhang et al., 2021). As relevance is a major aspect of dialogue coherence, we include these baselines for completeness. As both metrics are graph neural networks intended for larger train sets, we use checkpoints provided by the authors. GRADE is trained on DailyDialogue (Li et al., 2017), and DynaEval on Empathetic Dialogue (Rashkin et al., 2019). Both are trained with negative sampling, with GRADE constructing more challenging negative samples.

A summary of the authors' stated purpose for each metric can be found in the Appendix C.

4.1 Analysis

Table 1 makes it clear that the normalized probability and cosine similarity approaches do not generalize well across datasets. Although NORM-PROB excels on the P-DD dataset, it has weak performance on HUMOD and a significant *negative* cor-

relation on USR-TC. Likewise the FED metrics perform well on the FED data, but are negatively correlated on all other datasets. Consequently, we believe that the NORM-PROB and FED metrics are overfitted to their corresponding datasets. Similarly, although COS-FT 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.

Looking at the *dialogue coherence* metrics, DYNA-EVAL performs strongly on FED, and weakly on all other datasets. GRADE performs very strongly on HUMOD and P-DD (the latter, likely in part as it was trained on DailyDialogue), but is uncorrelated on USR-TC. Given that these metrics were not intended to measure relevance, uneven performance is to be expected as relevance and *dialogue coherence* will not always align.

The final baseline, NUP-BERT, is quite competitive, outperforming each of the other baselines on at least 2 of the datasets. Despite this, we can see that performance on HUMOD, USR-TC, and FED is still fairly weak. We can also observe that NUP-BERT has some sensitivity to the domain of the training data; fine-tuning on HUMOD data results in lower Spearman's correlation on USR-TC, and fine-tuning on USR-TC performs worse on the FED datasets. However, the amount of training data (TC vs TC-S) has little impact.

Overall, the results of Table 1 are concerning as they suggest that at least five current approaches generalize poorly across either dialogue models or domains. The absolute performance of all metrics studied vary considerably by dataset, and the relative performance of closely related metrics such as COS-FT and COS-NSP-BERT, or NUP-BERT with different training data, varies considerably between datasets. As a result, research into new dialogue relevance metrics is required. Furthermore, it is clear that the area's evaluation methodology must be updated to use various dialogue models in various different domains.

5 IDK: A 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 architecture is mostly unchanged, however the training procedure and the features used are altered.

First, based on the observation that the amount of training data has little impact, we freeze BERT features and do not fine-tune to the domain. Additionally, whereas the NUP-BERT baseline uses max-pooled BERT word embeddings, we use the pre-trained next sentence prediction (NSP) features: "(classification token) further processed by a Linear layer and a Tanh activation function [...] trained from the next sentence prediction (classification) objective during pre-training"[4].

Second, to improve generalization and reduce variation in training (particularly important as the practitioner typically has no 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 ($\lambda = 1$). Note that experiments with L2 regularization yielded similar validation set performance (see Appendix, Table 9).

Third, in place of random sampling we use a fixed negative sample, "i don't know". This allows us to train the model on less data.

Additionally, we simplify the model, using logistic regression in place of 2-class softmax. We train for 2 epochs using BCE 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, and batch size 6.

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, surpassing HUMOD baselines (Merdivan et al., 2020), and achieving parity with GRADE (0.61). Examples of the our metric's output on the HUMOD dataset, and a scatter plot of IDK vs human scores are in Appendices A and F, respectively.

Compared to NUP-BERT, our proposed metric provides strong improvement on the HUMOD dataset and equivalent or stronger performance on USR-TC and FED, at a cost of performance on P-DD. In particular, IDK (TC-S) performance on the FED datasets is considerably stronger than NUP-BERT (TC-S). 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,

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

we consider this tradeoff to be a net benefit.

Compared to GRADE in particular, we have reduced performance on P-DD, equivalent performance on HUMOD, and stronger performance on USR-TC and FED (in particular, correlation on the USR-TC dataset is non-zero). It is worth noting that, in general, our approach does not out-perform the 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.

Our metric is also preferable, as it is less sensitive to domain. To numerically demonstrate this, we measure the domain sensitivity of the evaluated metrics as the ratio of best Spearman's correlation to worst Spearman's correlation – this value should be positive (i.e., there is no dataset where the metric becomes negatively correlated), and as close to 1 as possible (i.e., there is no difference in performance). Looking at Table 9, we find IDK strongly outperforms all prior metrics, reducing this ratio by more than 37%-66% compared to the best baseline.

Prior Metric	Ratio
FED-CORRECT	-0.7
FED-RELEVANT	-0.7
NORM-PROB	-2.7
COS-NSP-BERT	-7.5
COS-FT	-13
GRADE	∞
DYNA-EVAL	∞
NUP-BERT (TC-S)	11.6
NUP-BERT (TC)	10.3
COS-MAX-BERT	6.7
NUP-BERT (H)	6.2
IDK (H)	**3.9**
IDK (TC-S)	**3.9**

Table 3: Ratio of best Spearman correlation to worst on all datasets for all metrics. Sorted in improving order.

5.1 Testing NSP feature dimensionality

As a followup experiment, we tested our assumption that only a fraction of the BERT-NSP features are needed. Plotting the weights learned by IDK on HUMOD, we found a skewed distribution with a small fraction of weights with magnitude above 0.01 (See Appendix, Figure 1). Hypothesizing that the largest weights correspond to the relevant dimensions, we modified the pretrained huggingface NSP BERT to zero all dimensions of the NSP feature, except for the 7 dimensions corresponding to the largest IDK HUMOD weights. We then evaluated NSP accuracy on three NLTK (Bird et al., 2009) corpora: Brown, Gutenburg, and Webtext. As expected, we found that reducing the dimensionality from 768 to 7 had no negative impact (see Appendix, Table 6). Again, note that the mask was created using IDK trained on HUMOD data, and the weights of BERT and the NSP prediction head were in no way changed. Therefore, it is clear that (at least on these datasets) over 99% of the BERT NSP feature dimensions can be safely discarded.

5.2 Ablation tests

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

Overall, it appears that the majority of the performance gains come from the combination of L1 regularization with pretrained BERT NSP features. The clearest observation is that L1 regularization is critical to good performance when using "i don't know" in place of random samples – otherwise, the model presumably overfits. Second, using "i don't know" in place of random samples has a mixed, but relatively minor effect. Thirdly, 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 a validation set of domain-specific 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.

The result that L1 regularization allows us to use "i don't know" in place of random negatives samples is quite interesting, as it seems to counter work in contrastive representation learning (Robinson et al., 2021), and dialogue quality evaluation (Lan et al., 2020) suggesting that "harder" negative examples are better. We believe that the reason for this apparent discrepancy is that *we are not performing feature learning*; the feature space is fixed, pretrained, BERT NSP. Furthermore, we've shown that this feature space is effectively 7 dimensional. As a result, we believe that the L1 regularization causes an effective projection to 7D. Consequently, as our model is low-capacity, "i don't know" is sufficient to find the separating hyperplane. Having said this, it is still unclear why we see *improved* performance on FED when training on HUMOD data. Comparing the histograms of learned weight magnitudes (see Appendix, Figure 2) we find that the ablated model has larger number of large weights – we speculate that the random

			HUMOD		USR-TC		P-DD		FED-Correctness		FED-Relevance	
Data	L1	idk	S	P	S	P	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)	*0.15 (0.00)	*0.23 (0.00)	*0.24 (0.00)	*0.29 (0.00)
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)	*0.14 (0.00)	*0.17 (0.01)	*0.21 (0.01)	*0.19 (0.02)
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)	0.09 (0.00)	*0.19 (0.01)	*0.17 (0.00)	*0.26 (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)	0.10 (0.02)	†0.11 (0.02)	†0.14 (0.02)	0.09 (0.03)
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)	*0.15 (0.00)	*0.23 (0.00)	*0.24 (0.00)	*0.29 (0.00)
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)	*0.14 (0.00)	*0.15 (0.01)	*0.21 (0.00)	*0.17 (0.01)
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)	†0.14 (0.01)	*0.21 (0.00)	*0.22 (0.01)	*0.29 (0.01)
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)	0.05 (0.06)	0.11 (0.03)	†0.12 (0.12)	†0.18 (0.04)

Table 4: Test correlation of various ablations of the proposed metric. The L1 column signifies whether L1 regularization is used ($\lambda = 1$), 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.

negative samples' variation in irrelevant aspects such as syntactic structure is responsible.

5.3 Additional Experiments

We repeated our IDK experiments with two different fixed negative samples; performance and domain sensitivity are generally comparable, although unexpectedly more sensitive to the choice of training data (see Appendix J). We also experimented with using the pretrained BERT NSP predictor as a measure of relevance, however performance is considerably worse on the longer-context FED dataset (see Appendix I). Finally, we observed that BCE loss encourages the model to always map "i don't know" to zero; yet, the relevance of "i don't know" varies by context. Unfortunately, experiments with a modified triplet loss did not yield improvements (see Appendix H).

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, there is a bidirectional flow of innovations. 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 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 evaluation have been motivated by dialogue relevance. There is a long history of evaluating dialogue models on specific aspects; 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, 2020b; Phy et al., 2020; Berlot-Attwell and Rudzicz, 2021). These approaches also share heritage with earlier ensemble measures of dialogue evaluation such as RUBER (Tao et al., 2018) – although in the case of RUBER, it combined a referenced and unreferenced metric rather than separate aspects.

Metrics of dialogue relevance and quality also share common problems such as the diversity of valid responses. Our findings that existing relevance metrics generalize poorly to new domains is consistent with previous findings about metrics of dialogue quality (Lowe, 2019; Yeh et al., 2021). Thus, our work suggests that this challenge extends to the subproblem of dialogue relevance as well.

At the same time, it must be remembered that measuring holistic dialogue quality is a very different task from measuring dialogue relevance – it is well established that aspects of dialogue such as fluency, and interestingness are major components of quality (Mehri and Eskenazi, 2020b,a), and these should have no impact on relevance.

With respect to prior work comparing relevance metrics, we are aware of only one tangential work. Yeh et al. (2021) performed a comparison of various metrics of dialogue *quality*; within this work they dedicated three paragraphs to a brief comparison of how these *quality* metrics performed at predicting various dialogue qualities, including rel-

evance. They reported results on only two of the datasets we used (P-DD and FED). Interestingly, the authors found that the FED metric performs well on P-DD (reporting a Spearman's correlation of 0.507), however our results demonstrate that the *components* of FED that are meant to measure relevance (i.e. FED-REL and FED-COR) are significantly *negatively* correlated with human relevance scores. Additionally, as Yeh et al. (2021) focus on quality, they do not compare performance between the two relevance datasets. Instead they compare performance on quality against performance on relevance, and use the discrepancy to conclude that measuring relevance alone (as done by NORM-PROB) is insufficient to determine quality. Although we agree that relevance alone is insufficient for dialogue quality evaluation, our work provides a richer understanding. Our finding that NORM-PROB performs poorly across a range of relevance datasets suggests that the poor performance of NORM-PROB in the quality-prediction task is also caused by the *poor relevance generalization* in addition to the insufficiency of relevance to measure overall quality.

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 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 is wise to collect a validation dataset of human-annotated relevance data for use in selecting a relevance metric. If this is not possible, then our metric, IDK, appears to be the best option – achieving both good correlation and the lowest domain sensitivity, even when trained on different domains. Furthermore, 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. If that is still too data intensive, then our results suggest that our metric is fairly agnostic to the domain of the training data; therefore 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 merited. The results are often counter-intuitive; our demonstration that 99% of the BERT NSP features can be safely discarded is just one striking example. Similarly, although our empirical results suggest that use of a single negative example generalizes across domains, there is no compelling theoretical reason why this should be so. More generally, all the metrics outlined are complex, dependent on large corpora, and created without ground truth annotations. 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 – 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 pre-trained BERT, or similar architectures), our aim is to provide a useful metric for rapid prototyping and hyperparameter search. Additionally, we 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.

8 Conclusion

Our work demonstrates that several existing metrics of dialogue relevance are problematic as their performance varies wildly between test-domains. We take a first step towards resolving this issue by proposing IDK: a simple metric that is less sensitive to test domain and trainable with minimal data. We reduce IDK's data requirements through the novel use of a fixed negative example, provide evidence that the underlying BERT NSP features are low-dimensional, and propose that this fact (combined with IDK's lack of feature learning) allows for the counter-intuitive use of a single negative example. Beyond this, we call for better evaluation of future relevance metrics, and thus release our code for processing four diverse, publicly available, relevance-annotated data sets.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

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 Arbor, 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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly.

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. In *Logic and Argumentation*, pages 477–485, Cham. Springer International Publishing.

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*.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

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*.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. PONE: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Trans. Inf. Syst.*, 39(1):7:1–7:37.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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*, pages 110–119, San Diego, California. Association for Computational Linguistics.

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.

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

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. 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.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems.

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.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5676–5689, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. 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 Example Evaluations

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 5: 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, the scores generated by our metric (IDK trained on HUMOD) are linearly shifted and re-scaled to 1-5.

B NSP Masking Experiment Results

The results of the NSP masking experiment are outlined in Table 6. Note that masking $> 99\%$ of the NSP feature had no impact on the pretrained model, and actually improved accuracy by 2.8% on the Webtext corpus.

C Exact objectives of prior metrics

In this section, we briefly outline the stated purpose of each of our relevance metrics evaluated:

Masked	Brown	Gutenburg	Webtext
	85.7%	75.3%	65.4%
✓	85.6%	**75.5%**	**68.2%**

Table 6: Next Sentence Prediction (NSP) performance on various NLTK (Bird et al., 2009) corpora using a pre-trained BERT and NSP head. When masked, we zero-out the 768-dim BERT NSP feature, leaving only the 7 dimensions corresponding to the largest magnitude weights in IDK (H) (i.e., we zero out $> 99\%$ of the feature vector).

- COS-FT: "In this work, given a dialogue history, we regard as a coherent response an utterance that is thematically correlated and naturally continuing from the previous turns, as well as lexically diverse." (Xu et al., 2018)
- NUP-BERT: "Maintains Context: Does the response serve as a valid continuation of the preceding conversation?" (Mehri and Eskenazi, 2020b)
- NORM-PROB: "context coherence of a dialogue: the meaningfulness of a response within the context of prior query" (Pang et al., 2020)
- FED-REL: "Is the response relevant to the conversation?" (Mehri and Eskenazi, 2020a)
- FED-COR: "Is the response correct or was there a misunderstanding of the conversation? [...] No one has specifically used Correct, however its meaning is often encapsulated in Relevant." (Mehri and Eskenazi, 2020a)

We also outline the stated purpose of the *dialogue coherence* metrics evaluated:

- GRADE: "Coherence, what makes dialogue utterances unified rather than a random group of sentences" (Huang et al., 2020)
- DYNA-EVAL: "dialogue coherence: considers whether a piece of text is in a consistent and logical manner, as opposed to a random collection of sentences" (Zhang et al., 2021)

D Details for Prior work

D.1 NORM-PROB

Pang et al. (2020) relied on a pretrained 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 5th percentile of the distribution of $\frac{1}{|r|} \log P(r \mid q)$ over the examples being evaluated.

E Learned HUMOD-IDK Weights

Figure 1 depicts the distribution of weight-magnitudes learned by IDK on the HUMOD training set. Notably, there is a very small subset of weights which is an order of magnitude larger than the others. Figure 2 demonstrates that the use of random sampling in place of "i don't know" when training on the HUMOD dataset causes a larger number of large weights.

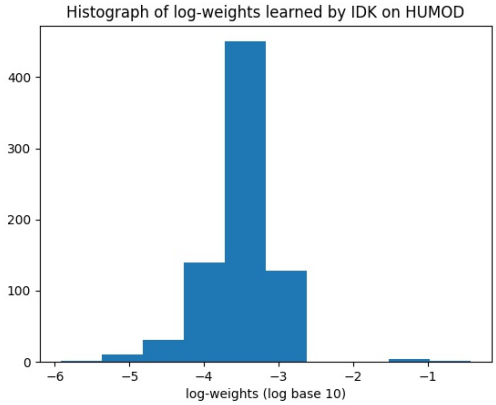

Figure 1: Histogram of log weight magnitudes learned by IDK on HUMOD. Note the small number of weights that are an order of magnitude larger.

F Scatter Plots

Figures 3, 4, 5, 6, and 7 illustrate IDK vs human scores of relevance, where the IDK training data is HUMOD. A regression line is fitted to highlight the trend.

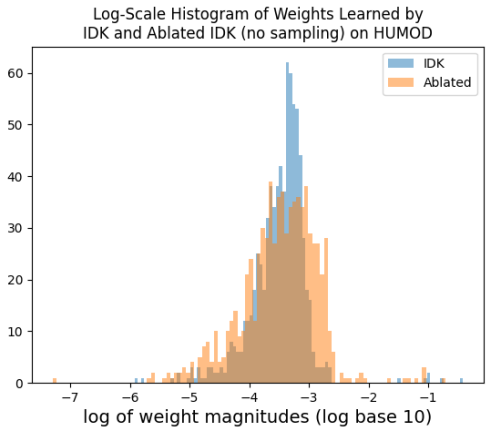

Figure 2: Histogram of log weight magnitudes learned by IDK and Ablated IDK on HUMOD. The specific ablation is the use of random negative samples in place of "i don't know". Note that Ablated IDK has a larger number of large weights than normal IDK.

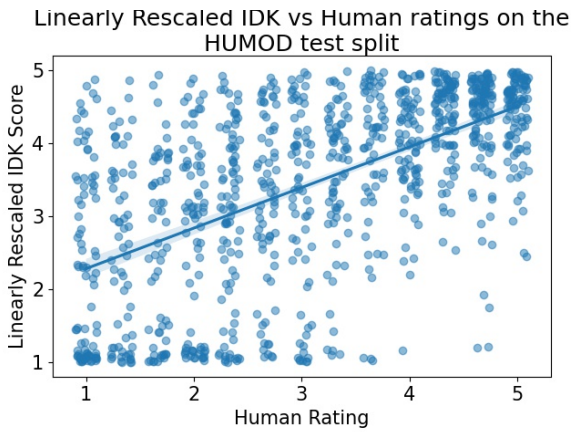

Figure 3: IDK scores, linearly re-scaled to the range 1-5, versus human scores of relevance, on the HUMOD test set.

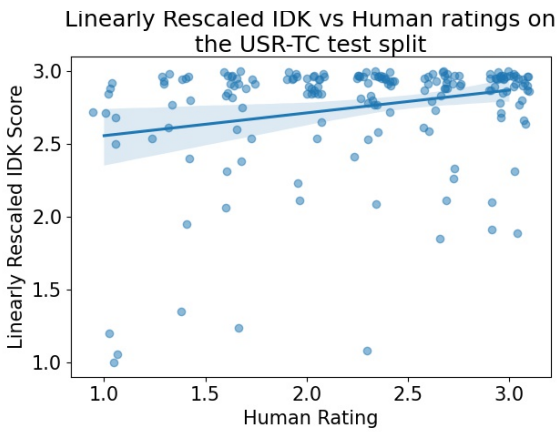

Figure 4: IDK scores, linearly re-scaled to the range 1-3, versus human scores of relevance, on the USR-TC test set.

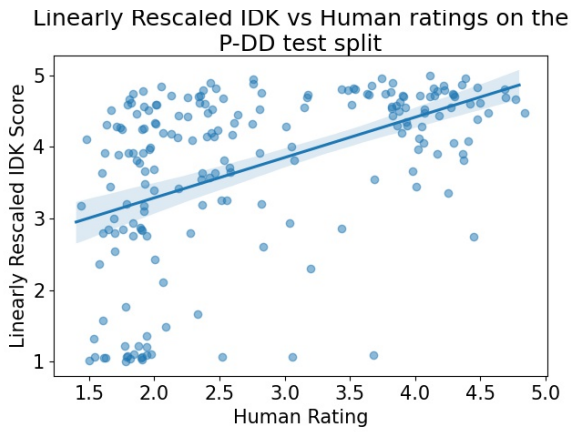

Figure 5: IDK scores, linearly re-scaled to the range 1-5, versus human scores of relevance, on the P-DD test set.

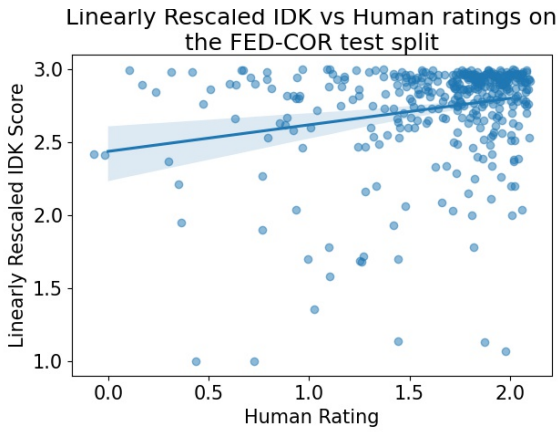

Figure 6: IDK scores, linearly re-scaled to the range 1-3, versus human scores of relevance, on the FED-CORRECT test set.

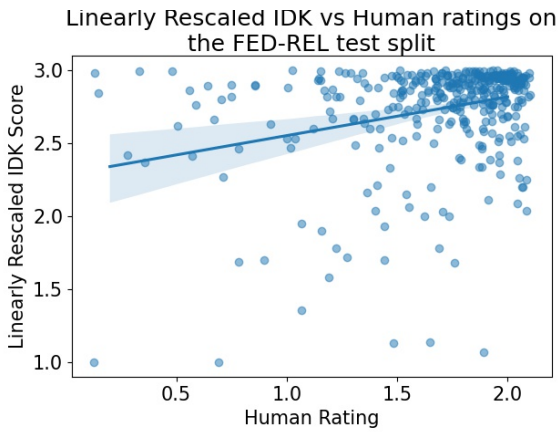

Figure 7: IDK scores, linearly re-scaled to the range 1-3, versus human scores of relevance, on the FED-RELEVANT test set.

G Performance on validation data split

Correlations of the models on the validation set are outlined in Table 7 for prior metrics, and in Table 9 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
FED-CORRECT	*−0.10	*−0.09	−0.14	−0.15
FED-RELEVANT	*−0.10	*−0.09	−0.14	−0.16
GRADE	*0.64	*0.64	0.02	0.00
DYNA-EVAL	*0.14	*0.15	−0.05	−0.06
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 7: 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. NOTE: USR scores are human only for COS-FT, NORM-PROB and NUP-BERT

			HUMOD		USR-TC		P-DD		FED-Correctness		FED-Relevance	
Data	L1	idk	S	P	S	P	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)	†0.13 (0.02)	*0.21 (0.01)	*0.21 (0.01)	*0.30 (0.00)
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)	0.10 (0.02)	†0.11 (0.02)	†0.14 (0.02)	0.09 (0.03)
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)	0.11 (0.02)	†0.18 (0.06)	*0.20 (0.02)	*0.25 (0.08)
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)	†0.14 (0.00)	*0.20 (0.02)	*0.22 (0.00)	*0.27 (0.02)
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)	0.10 (0.04)	†0.10 (0.14)	†0.17 (0.07)	†0.14 (0.21)
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)	†0.13 (0.01)	*0.15 (0.01)	*0.19 (0.02)	*0.17 (0.02)
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)	†-0.05 (0.18)	*-0.07 (0.26)	*-0.08 (0.27)	*-0.09 (0.35)
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)	0.02 (0.10)	†0.08 (0.07)	0.00 (0.10)	†0.12 (0.10)

Table 8: Repeat of ablation experiments, instead using modified triplet loss ($m = 0.4$) in place of BCE. 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).

Name	HUMOD Spear	HUMOD Pear	TC Spear	TC Pear
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)
H_IDK_L1	*0.56 (0.01)	*0.53 (0.02)	*0.45 (0.03)	*0.44 (0.02)
H_IDK_L2	*0.55 (0.00)	*0.55 (0.01)	*0.44 (0.00)	*0.44 (0.00)
H_Rand3750_L1	*0.42 (0.22)	*0.40 (0.20)	*0.44 (0.00)	*0.45 (0.01)
H_Rand3750_L2	*0.56 (0.00)	*0.55 (0.01)	*0.45 (0.00)	*0.44 (0.02)
H_Rand3750_bce_L1	*0.58 (0.00)	*0.58 (0.00)	*0.45 (0.00)	*0.46 (0.00)
H_Rand3750_bce_L2	*0.57 (0.00)	*0.56 (0.00)	*0.45 (0.00)	*0.42 (0.00)
H_IDK_bce_L1	*0.57 (0.00)	*0.56 (0.00)	*0.42 (0.01)	*0.41 (0.00)
H_IDK_bce_L2	*0.50 (0.01)	*0.51 (0.01)	*0.39 (0.00)	*0.42 (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)
TC-S_IDK_L1	*0.29 (0.43)	*0.23 (0.53)	*0.39 (0.07)	*0.41 (0.07)
TC-S_IDK_L2	*0.54 (0.01)	*0.55 (0.01)	*0.43 (0.01)	*0.44 (0.00)
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_L2	*0.47 (0.02)	*0.48 (0.01)	*0.41 (0.00)	*0.39 (0.01)
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_L2	†-0.33 (0.27)	†-0.32 (0.26)	*-0.45 (0.02)	*-0.43 (0.02)
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_L2	*0.04 (0.55)	*0.09 (0.56)	†-0.26 (0.27)	†-0.23 (0.31)
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 9: 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 are "i don't know" or random. If bce is present, then BCE was used as the loss, otherwise our modified triplet loss is used. If L1 or L2 is present, then L1 or L2 regularization with $\lambda = 1$ is used respectively, 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 one trial was significantly different from zero at $p < 0.01$. Note that L1 and L2 regularization have similar effects, with the exception of worse performance between TC-S_Rand2750_bce_L1 and TC-S_Rand2750_bce_L2; we suspect this could be overcome with hyperparameter tuning.

	HUMOD		USR-TC		P-DD		FED-Correctness		FED-Relevance	
Prior Metric	S	P	S	P	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** (0.04)	*0.54 (0.02)	†0.14 (0.04)	*0.21 (0.03)	*0.22 (0.01)	***0.30** (0.01)
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)	0.05 (0.04)	0.12 (0.01)	†0.16 (0.04)	*0.21 (0.01)
NUP-BERT (TC)	*0.30 (0.01)	*0.38 (0.00)	0.16 (0.02)	*0.21 (0.02)	***0.62** (0.05)	***0.58** (0.04)	0.06 (0.01)	†0.12 (0.02)	*0.18 (0.02)	*0.23 (0.01)
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)	*0.15 (0.00)	***0.23** (0.00)	***0.24** (0.00)	*0.29 (0.00)
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)	*0.15 (0.00)	***0.23** (0.00)	***0.24** (0.00)	*0.29 (0.00)
IDK-ICS (H)	*0.55 (0.01)	*0.53 (0.00)	***0.25** (0.01)	***0.27** (0.00)	*0.44 (0.01)	*0.39 (0.00)	***0.16** (0.00)	*0.22 (0.00)	*0.22 (0.00)	***0.30** (0.00)
IDK-ICS (TC-S)	*0.58 (0.00)	*0.47 (0.00)	0.17 (0.00)	***0.27** (0.00)	*0.52 (0.00)	*0.36 (0.00)	*0.14 (0.00)	*0.16 (0.00)	*0.22 (0.00)	*0.24 (0.00)
IDK-OK (H)	*0.58 (0.00)	***0.59** (0.00)	0.15 (0.00)	*0.23 (0.00)	*0.49 (0.00)	*0.47 (0.00)	0.11 (0.00)	*0.19 (0.00)	*0.19 (0.00)	*0.26 (0.00)
IDK-OK (TC-S)	***0.59** (0.00)	***0.59** (0.00)	0.18 (0.00)	*0.24 (0.00)	*0.52 (0.00)	*0.46 (0.00)	*0.15 (0.00)	***0.23** (0.00)	*0.23 (0.00)	*0.29 (0.00)
BERT NSP	***0.59**	*0.40	0.17	*0.25	*0.53	*0.31	0.12	0.10	*0.21	*0.18

Table 10: Comparison of our proposed metric (IDK) against the pretrained BERT NSP predictor on the test set. We also trained IDK with different fixed negative examples, "i couldn't say" (IDK-ICS) and "i'm ok." (IDK-OK). Note BERT NSP tends to have comparable Spearman's performance and worse Pearson's correlation. The only exception is FED where BERT NSP has inferior performance. In general, IDK with different fixed negative samples outperforms NUP-BERT, and is less sensitive to training data, although not to the same extent as baseline IDK.

H Additional Experiments: Triplet Loss

An intuitive limitation of using "i don't know" as a negative example with BCE 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 interrupted 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, $\mathcal{L}(c, r) = -\log(1 + m - f_t(c, r))$ where $f_t(c, r) = \max(y(c, r) - y(c, r') + m, 0)$.

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 8) suggest equivalence, at best. Often, this loss performs equivalently to 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 BCE 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.

I Additional Experiments: BERT NSP

As a followup experiment we compared IDK against directly using the pretrained BERT NSP predictor. In general, Spearman's correlation was comparable on all datasets *except for FED*, and Pearson's correlation was degraded. Performance on FED was inferior to IDK. We speculate that the reason for this is that the FED datasets has longer contexts, which is problematic for the NSP predictor as it was trained with sentences rather than utterances. Results are summarized in Table 10.

J Additional Experiments: IDK with other fixed negative samples

As a followup experiment we trained IDK using two different fixed negative samples: "i couldn't say" (simply chosen as a synonym for "i don't know"), and "i'm ok." (chosen as an example of a generic response from Li et al. (2016)). Results are reported in Table 10; in general we still see an performance improvement over NUP-BERT, and in some cases we exceed the performance of baseline IDK. We also see that performance remains consistent between runs, maintaining a lower standard deviation than NUP-BERT.

However, it is also clear that changing the fixed negative sample has some unexpected consquences: specifically, we see variation based on training data that is not observed when using "i don't know" as the fixed negative sample (although the variation due to training data appears to be less than NUP-BERT).

We retain the reduced sensitivity to test set. Specically, our ratios of best-to-worst Spearman's correlation are 3.44 for IDK-ICS (H), 4.14 for IDK-ICS (TC-S), 5.27 for IDK-OK (H), and 3.93; most are very close to the baseline IDK ratio of 3.9, and all are an improvement on the best prior work; 6.2 on NUP-BERT (H) – it is worth noting that NUP-BERT (TC-S) attains a ratio of 11.6, considerably worse than when trained on HUMOD data.