Dialogue Response Ranking Training with Large-Scale Human Feedback Data

Xiang Gao **Yizhe Zhang** **Michel Galley**
Chris Brockett **Bill Dolan**
Microsoft Research, Redmond, WA, USA
{xiag,yizzhang,mgalley,chrisbkt,billdol}@microsoft.com

Abstract

Existing open-domain dialog models are generally trained to minimize the perplexity of target human responses. However, some human replies are more engaging than others, spawning more followup interactions. Current conversational models are increasingly capable of producing turns that are context-relevant, but in order to produce compelling agents, these models need to be able to predict and optimize for turns that are genuinely engaging. We leverage social media feedback data (number of replies and upvotes) to build a large-scale training dataset for feedback prediction. To alleviate possible distortion between the feedback and engagingness, we convert the ranking problem to a comparison of response pairs which involve few confounding factors. We trained DIALOGRPT, a set of GPT-2 based models on 133M pairs of human feedback data and the resulting ranker outperformed several baselines. Particularly, our ranker outperforms the conventional dialog perplexity baseline with a large margin on predicting Reddit feedback. We finally combine the feedback prediction models and a human-like scoring model to rank the machine-generated dialog responses. Crowd-sourced human evaluation shows that our ranking method correlates better with real human preferences than baseline models.[1]

1 Introduction

Conversing freely in natural language is one of the greatest challenges of artificial intelligence. End-to-end open-domain dialog systems have become increasingly powerful, with advanced model architectures and large-scale training (Zhang et al., 2019b; Adiwardana et al., 2020; Roller et al., 2020; Li et al., 2020). In some settings, human annotators cannot reliably distinguish between human- and machine-generated responses. Though surprisingly effective, the training objective for these models is conceptually simple: minimizing the perplexity of a reference response for a given context.

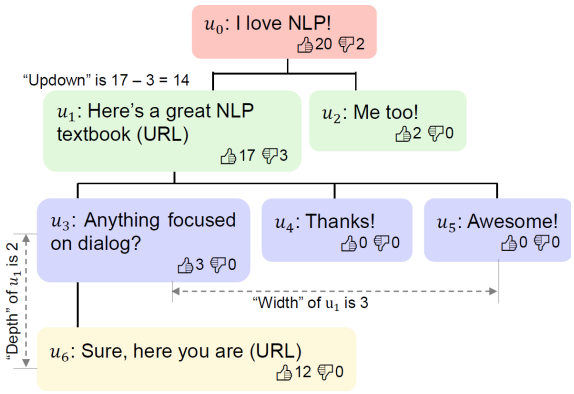

Figure 1: For many online communities, posts and comments have a tree structure and user can upvote or downvote each node individually. This allows us to define measures (e.g. Width, Depth, and Updown) of human feedback and build a large-scale training dataset for response quality prediction.

However, a meaningful evaluation of response generation must take into account more than whether a generated turn is relevant in context, or whether it "sounds human." Conventional neural conversation models often generate trivial or bland responses (Li et al., 2016; Zhao et al., 2017) that are relevant to context but are not engaging. Even human responses can vary dramatically in terms of tonal appropriateness and whether they are interesting enough to prompt a rich listener reaction. A successful dialog turn must be proactive, engaging, and consistent with social norms (Grice, 1975, 1989).

In this work, we move beyond simple prediction of response relevance, augmenting this with a prediction of how likely a response is to elicit a

[1]Dataset and models open-sourced on https://github.com/golsun/DialogRPT

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 386–395,
November 16–20, 2020.

positive reaction from an interlocutor. By incorporating a measure of engagingness into the response generation ranking algorithm, we hope to improve the overall behavior of data-driven conversational agents.

Existing methods are suboptimal for this ranking task. Conventional perplexity based ranking methods (Li et al., 2016; Vijayakumar et al., 2016) focus only on context-hypothesis relevancy. Online conversational systems such as XiaoIce (Zhou et al., 2018) employ a manually-designed set of features to rank hypotheses, but the design of these rankers is not directly based on real-world human preferences or feedback in an end-to-end fashion. Large-scale training data is necessary because of the one-to-many nature of dialog and the scope and complexity of human conversation. However, labeling conversations at scale is too expensive and time-consuming for this purpose. Labeling the "engagingness" of a response is not something a single annotator can do; the task requires something more like a large-scale, collective vote. And yet there is no obvious automated substitute for this kind of human labeling. Conventional quality measurements such as reference-based similarity (Papineni et al., 2002) or lexical diversity (Li et al., 2016; Zhang et al., 2018b) capture only limited aspects of response quality, and are not strongly predictive of human reactions: simply because a response is different from others does not necesarily mean that it will be perceived as "bad".

Our solution involves leveraging existing human feedback data (e.g., number of replies and likes) from online social communities. While there is work in the field of social media on feedback prediction (Sparling and Sen, 2011; Stoddard, 2015; Glenski and Weninger, 2017), it has not previously been applied to dialog systems and response generation. As illustrated in Figure 1, each comment has its own number of replies and upvotes (termed as "Likes" in some communities). These can be used as engagingness labels after careful normalization and formulation. There exist billions of online threads available and the number is growing fast, thus making it possible to build a large-scale training dataset. However, the relation between feedback and quality may be distorted due to social influence and other confounding factors (Salganik et al., 2006).

In order to ameliorate this problem, we propose a contrastive formulation, shifting from ranking to pairwise classification. Using a dataset of 133M pairs of human comments and their associated number of replies or up-/downvotes, we train a set of large-scale transformer-based feedback ranking models which outperform several baselines. In particular, dialog perplexity shows little predictive power of human feedback. We also show that a classifier trained on human-vs-artificial data can achieve good zero-shot relevancy prediction accuracy. Finally, we describe an ensemble model that is capable of merging the predictive powers of all these models, tuned using human calibration. Human evaluation shows that our ranking method outperforms the baselines in terms of correlation with actual human preferences.

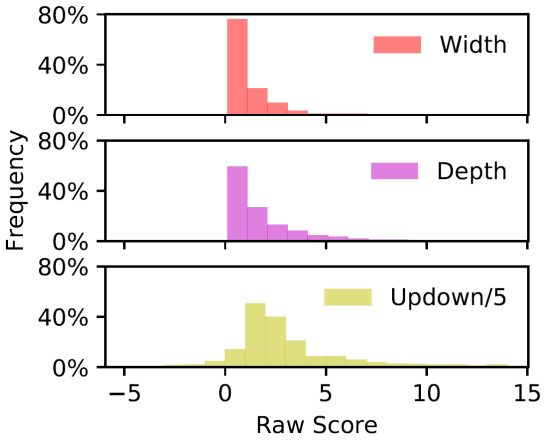

Figure 2: The long-tailed distribution of the raw scores of feedback of Reddit.com.

2 Human Feedback

Many social media platforms, such as Reddit, Twitter, and Facebook allow users to reply or upvote contents, leveraging that feedback to make decisions about what content to display, highlight, and hide. These collective ratings are treated as a proxy for content engagingness. In this section we discuss a few metrics of user vote data, along with some of the issues posed by its use.

	Width	Depth	Updown
Width	1	0.8592	0.3491
Depth	0.8592	1	0.3257
Updown	0.3491	0.3257	1

Table 1: Spearman's ρ between different measurements of human feedback. Darker cell color indicates higher correlation.

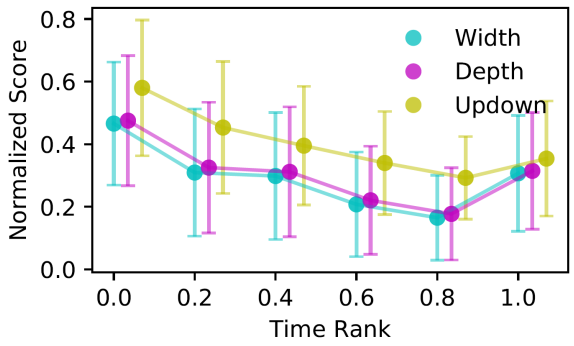

Figure 3: The dependence of feedback on created time of the comments of the same parent node (i.e. the same context) of Reddit.com. Error bars show standard deviation.

2.1 Feedback metrics

As illustrated by Figure 1, posts and comments typically form a tree structure. Each comment branching from the root may have its own comment children. We consider the path from the root to the parent node of a comment to be its context c, and the comment as a reply r. For each dialog (c, r), we consider the following feedback: **Width**, the number of direct replies to r; **Depth**, the maximum length of the dialog after this turn; and **Updown**, the number of upvotes minus the number of downvotes. For example, given the context $c = u_0$, the reply u_1 gets three direct replies u_3, u_4, u_5 and the Width is thus 3. u_3 continues the dialog with one more turn u_6, thus the depth is 2. u_1 got 17 upvotes and 3 downvotes so its Updown is 14. In contrast, u_2 is for the same context, but its Width and Depth is only 0, and Updown is 2.

Though focused on different dimensions, both Width and Depth can be seen as measures of the number of replies, and are therefore often closely correlated, as shown in Table 1 using Reddit as an example. They are less correlated with Updown. Presumably, contributors may feel that an upvote is enough to express their agreement or appreciation, and so do not post a full reply.

2.2 Feedback and Engagingness

The feedback metrics defined above cannot be directly used as a measure of reply engagingness. Stoddard (2015) shows that while popularity, measured by Updown, generally increases with quality, posts of similar quality can exhibit very different upvote counts. This variability can be traced to several different factors. As illustrated in Figure 2, the distribution of feedback is long-tailed, with a small fraction of threads receiving most of the replies and likes. Additionally, the popularity of the specific subreddit in which a comment occurs further confounds things: a relatively uninteresting comment in a very popular thread may get more feedback than an interesting comment in a less trafficked subreddit. Feedback volume is also heavily dependent on the timing of a comment relative to other comments, with replies that come early in a thread being more likely to attract replies or likes. This is shown in Figure 3. This may be tied to other factors such as social influence and disparities in comment visibility causing distortions in the relationship between comment engagingness and popularity (Salganik et al., 2006; Salganik and Watts, 2008; Gilbert, 2013). These findings imply that careful formulation and normalization should be applied before using feedback data as a training signal. We present our approach to this in Section 3.1.

2.3 Tasks

Given a context and a list of responses, we consider the task of predicting a ranking based on the feedback they received, as measured by these three separate metrics: (1) Width, (2) Depth, and (3) Updown. The gold label and training data is available for human response ranking, but in order to make this applicable to machine generated responses, we introduce another task: (4) human-vs-fake, which measures how human-like the response is. We consider two modes of fake examples: random human responses and machine generated responses. We will introduce an ensemble method in Section 3.2 for this last task.

3 The DIALOGRPT Method

In this section we introduce Dialog Ranking Pretrained Transformers (DIALOGRPT).

3.1 Problem Formulation

A Contrastive Learning approach. Given the confounding factors affecting feedback mentioned above, we train the model on *pairs* of samples (c, r^+) and (c, r^-), rather than fitting it to score each dialog *individually*. This follows the Contrastive Learning approach (see Section 5 for a brief review). The model is trained to predict a higher score for the positive sample r^+ (i.e. the response with more feedback) compared to the negative sample r^-. Besides (1) only comparing replies of the

same context, we use the following criteria to construct pairs that minimize the effect of confounding factors: (2) the sequence of two replies, r^+ and r^-, must have been created within a brief time window (no more than one hour), and (3) the feedback score of r^+ must exceed that of r^- by a specified threshold in order to make the label less noisy. Due to the long-tailed distribution, we consider both an absolute-valued threshold and a percentage ranking threshold. Furthermore, if a reply has more downvotes than upvotes, it will not be considered as a positive sample, but can be used as a negative sample.

Training objective. The model should be able to output a score at testing time for a hypothesis r for a given context c. At training time, as formulated in Section 3.1, given two hypotheses for a context, the model should be able to identify which one has more feedback. To connect these two requirements, the model outputs a scalar h,

$$h(c, r) = \text{DIALOGRPT}(c, r) \tag{1}$$

At inference time, we compute the score $s(r|c)$

$$s(r|c) = \text{Sigmoid}(h(c, r)) \tag{2}$$

For training, the loss is designed to simultaneously maximize the positive sample score and minimize the negative sample score:

$$\mathcal{L} = -\sum_{i \in \text{batch}} \log \frac{e^{h(c_i, r_i^+)}}{e^{h(c_i, r_i^+)} + e^{h(c_i, r_i^-)}} \tag{3}$$

This can be interpreted as the cross entropy between the target distribution $\{P(r^+) = 1, P(r^-) = 0\}$ and the predicted distribution in Softmax form. Note the contrastive form is crucial, given that a loss function only maximizing $s(r^+|c)$ usually leads to a collapsed solution (Hadsell et al., 2006).

3.2 Model ensemble

For machine generation. The machine generation is required to be both *human-like* and *preferred by human*. To rank the machine generations, we factorize the probability of a joint distribution as follows:

$$\begin{aligned} &P(r = \text{preferred, human-like}|c) \\ =&P(r = \text{preferred}|r = \text{human-like}, c) \cdot \\ &P(r = \text{human-like}|c) \end{aligned} \tag{4}$$

We estimate the first term with the models trained on a human-vs-human ranker on each feedback metric $K \in \{\text{Width, Depth, Updown}\}$

$$P(r = \text{preferred}_K, \text{human-like}|c) \triangleq s_K(r|c) \tag{5}$$

We denote the term $P(r = \text{human-like}|c)$ as $\pi_0(r|c)$, and build a classifier to predict how human-like a response is (see Section 3.3 for details).

$$P(r = \text{human-like}|c) \triangleq \pi_0(r|c) \tag{6}$$

Both $\pi_0(r|c)$ and $s_K(r|c)$ are scores defined in Eq. 2 interpreted as probability.

For overall preference. In case only a simple human preference matters (instead of separate Width, Depth, Updown metrics), we assume that a linear combination exists

$$s_{\text{Prefer}}(r|c) \triangleq \pi_0(r|c) \sum_K w_K s_K(r|c) \tag{7}$$

Human calibration. To estimate the correlation between the feedback score and human response preference, we present pairs of responses for the same context to a set of human annotators, asking them to select the response they would prefer to send or receive. The annotation is conducted for machine-vs.-machine comparisons on 1K pairs, and with 5 individual judges for each pair. Through this controlled setup, we reduce confounding factors, such as social influence and disparities in visibility, that might exist even within the contrastive problem formulation.

The results are used as a proxy for $s_{\text{Prefer}}(r|c)$, and can be used to estimate w_K for the test set, though the optimal value may depend on the test set and the instructions the human annotators were given. Note that the freedom of the system is now limited to a handful of hyper-parameters, limiting the need for large-scale human labeling to learn the model parameters.

3.3 Implementation details

Model and training. Our model is a 12-layer transformer model based on GPT-2 (Radford et al., 2019) architecture, and initialized with DialoGPT-medium model weights (Zhang et al., 2019b). DialoGPT is a large-scale dialog response generation model, pre-trained on 147M Reddit conversations. We use a linear layer to convert the final layer transformer output at the last token time step to a scalar

h. The parameters of the transformers and this output layer are trained simultaneously.

Each model has 354.8M parameters, and is trained on an Nvidia Tesla V100 4-core GPU with batch size 256 at an average training speed of 0.33 M pairs of samples per hour. Each model took around 70 hours to converge (until validation loss on a fixed set of 1024 samples ceased to improve).

Model	Trained on	Dataset size (M)
Human feedback $s_K(r\|c)$	Human vs. Human	
	- Width	22.3
	- Depth	25.1
	- Updown	40.7
Human-like $\pi_0(r\|c)$	Human vs. Fake	
	- Rand	40.7
	- Generated	5.3

Table 2: Summary of models and training data of different tasks, size in millions (M) of pairs

Data construction. Following the contrastive learning approach introduced in Section 3.1, we constructed a 133M-pair training set using Reddit data from 2011-2012, as shown in Table 2. For each task, we sampled 1024 validation pairs from the 2012 data and 5K test pairs from the 2013 data. The train, validation and test data do not share any Reddit posts.

For the human-like (i.e. human-vs-fake) task, we consider two representative negative modes: retrieval and generative dialog model generation. For the former we simply construct negative examples by randomly sampling from the training data. For the latter we use DialoGPT with top-k decoding. Since DialoGPT is able to produce human-like responses in certain evaluation settings, we select only 5.3 M highly-rated human response as positive examples, instead of using all human responses. Note that our method can be extended to include other negative modes such as perturbations and excessive repetition, similar to the synthetic example creation using BLEURT (Sellam et al., 2020).

3.4 Baselines

We consider the following baselines:

Dialog perplexity (ppl.) This metric is calculated for both the forward model (i.e., predict the response from the context) and the reverse model (i.e. predict the context from the response). This ranking method was proposed by Li et al. (2016) and formulated to maximize mutual information (MMI) between the response and context. We use DialoGPT and its reverse model to compute ppl.

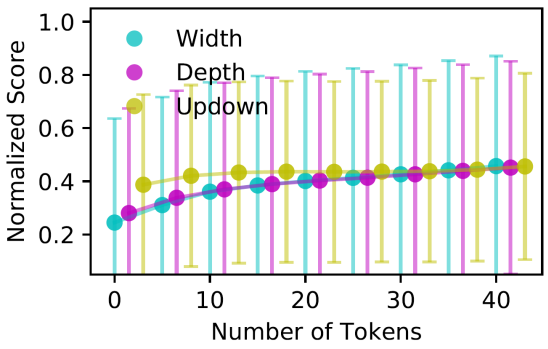

Figure 4: The dependence of feedback on the length for the comments of the same parent node (i.e. the same context). Error bars show standard deviation.

BM25 This classic metric measures keywords similarity (Robertson and Zaragoza, 2009). We use the inner product of the context BM25 vector and candidate response BM25 vector to rank candidates, similar to (Henderson et al., 2019a).

ConveRT (Henderson et al., 2019b) is a transformer-based model pretrained on Reddit data. It encodes context and candidate as vectors and compute their inner product as similarity used for ranking, achieved the existing state-of-the-art performance on several response matching test sets[2].

Bag of words (BoW) For each word, an average of rank-normalized feedback score[3] is calculated for replies that contain this word. This is the score for this word. Due to the long-tailed distribution of the absolute value of feedback items, we normalize them as the percentage ranking for their context. Then we use the average of the scores of the words in a response as the score of this response.

Length As shown in Figure 4, feedback rank weakly correlates with response length. We therefore use the average value of responses of the same length in training data as the predicted score for a hypothesis.

BoW and Length baselines are are intended to capture information about lexical patterns of hu-

[2] https://github.com/PolyAI-LDN/conversational-datasets/blob/master/BENCHMARKS.md

[3] defined as $1 - i/m$, where i is the feedback rank of this reply for the given context, and m is the number of replies of this context

man feedback in the data and provide a preliminary analysis.

4 Results

4.1 Predicting Human Feedback

Preliminary analysis We first consider findings from the bag of words baseline. As shown in Table 4, responses that receive fewer replies or upvotes tend to be less contentful (e.g. *lol*, *awesome*, *wow*, *nice*). In contrast, comments that attract more feedback are typically different in character: for instance, questions (indicated by *?*, *why*, *how*, *what*, *who*) often lead to longer conversation (greater Depth). Comments targeting a broad audience (labeled by *anyone*, *guys*), tend to receive more direct replies (greater Width) than those aimed at a specific set of people.

A similar pattern is captured by DIALOGRPT, as shown in Table 3. Given the context *I love NLP!*, the relatively bland response *Me too!* gets the lowest scores for all three feedback measures. Higher scores are obtained for Response *B*, where a justification is provided for the agreement (*useful*, *powerful*). Response *C* gets the highest Depth score, as it invites a discussion about how NLP works, something that is unlikely to be completed in one or two turns. Response *D*, in contrast, can be answered in fewer turns but with potentially many valid answers, which explains its high Width score. Finally, Response *E* receives the highest Updown score, probably because the model predicts that many people will upvote it to express gratitude for the useful resource pointer it provides (*textbook*). Removing the word *(URL)* from Response *E* causes the score to drop only slightly, indicating that the model is not simply sensitive to the post containing a web link.

Ranker evaluation We evaluate ranker performance using two metrics. First, we use pairwise accuracy, which measures accuracy in selecting the positive sample from a positive (more feedback) and negative (less feedback) pair for the same context. This is consistent with the training objective. Second, since the models will be used to rank hypotheses, we are also interested in the correlation between the model scorer rank and the the gold label rank. We measure this correlation using Spearman's ρ.

As shown in Table 5, DIALOGRPT shows the highest test performance on both measurements[4] Reverse dialog perplexity generally performs better than forward dialog perplexity. However, as it is not trained with feedback labels, a simple BoW baseline outperforms the dialog models in this task.

We also evaluated performance on feedback data that the model had not been trained on, as shown in Table 6.The model trained on Width data can perform reasonably well on Depth prediction, and vice versa, consistent with the high correlation between their labels as shown in Table 1. The Updown label is less correlated with these, and so the model trained on Updown data performs poorly on Width and Depth data. This is in keeping with the complementary relationship between these models.

4.2 Human-like Classification

Human-vs-Rand We first evaluate performance on the task of selecting the gold response from a set of random distractor responses. For each context, we randomly select n distractors. Performance is evaluated using Hits@k, which is the ratio of the number of gold responses in the top-k ranked hypotheses. Here, k is equal to the number of gold responses. Although DIALOGRPT is trained solely on Human-vs-Rand Reddit data, we show in Table 7 that it performs well even when compared to baseline models on other data sources: DailyDialog (Li et al., 2017) and Twitter[5] PersonaChat[6] (Zhang et al., 2018a). Such *zero-shot* performance indicate that the model generalize reasonably well on unseen datasets.

For the Reddit dataset, which has multiple gold replies, we also compare our method with reference-based similarity measurements, [7] including BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019a), and BLEURT (Sellam et al., 2020). These metrics are not applicable on-the-fly, since references are not available, but they are commonly used as offline measures of dialog system quality. As shown in Table 7, although BLEU, BERTScore, and BLEURT take advantage of reference, which is unknown to DIALOGRPT, DIALO-

[4] Similar results are observed for the validation set.

[5] https://github.com/Marsan-Ma/chat_corpus/

[6] The performance of IR Baseline, Starspace, and KV Profile Memory for PersonaChat are following Zhang et al. (2018a).

[7] Following Galley et al. (2018), for a gold hypothesis, we only use other $k-1$ gold hypotheses as references to avoid a similarity of 1. For each distractor response, we randomly pick $k-1$ references from k gold hypotheses.

Context: I love NLP!				
Response:		Width	Depth	Updown
A	Me too!	0.033	0.043	0.171
B	It's super useful and more and more powerful!	0.054	0.164	0.296
C	Can you tell me how it works?	0.644	**0.696**	0.348
D	Can anyone recommend a nice review paper?	**0.687**	0.562	0.332
E	Here's a free textbook (URL) in case anyone needs it.	0.319	0.409	**0.612**

Table 3: Predicted feedback scores of several example responses given the same context.

Width	r^+	url, anyone, else, who, does, why, guys, seriously, everyone
	r^-	oh, amazing, damn, thanks, wow, nice, !, awesome, lol, upvote
Depth	r^+	?, why, does, how, anyone, isn't, any, what, who,
	r^-	great, nice, amazing, damn, lol, !, awesome, thank, upvote
Updown	r^+	url, our, picture, everyone, hey, part, years, into, will, we
	r^-	maybe, though, awesome, comment, funny, wow, came, upvote, lol

Table 4: Bag of words analysis. If on average the comments containing a certain word get more feedback, we list this word in the r^+ row. If they get less feedback, this word is listed in r^- row.

	Method	Pairwise accuracy	Spearman ρ
Width	Dialog ppl.	0.513	-0.009
	Reverse dialog ppl.	0.571	0.099
	Length baseline	0.595	0.229
	BoW baseline	0.596	0.234
	DIALOGRPT	**0.752**	**0.357**
Depth	Dialog ppl.	0.508	-0.004
	Reverse dialog ppl.	0.557	0.063
	Length baseline	0.543	0.134
	BoW baseline	0.584	0.187
	DIALOGRPT	**0.695**	**0.317**
Updown	Dialog ppl.	0.488	0.003
	Reverse dialog ppl.	0.560	0.076
	Length baseline	0.531	0.063
	BoW baseline	0.571	0.134
	DIALOGRPT	**0.683**	**0.295**

Table 5: Performance on test set ranking gold responses, measured by pairwise accuracy and Spearman's ρ.

GRPTshows higher accuracy measured by Hits@k.

Human-vs-Generated We evaluate the model's ability to discriminate between human and generated responses. As shown in Table 6, a model trained only on human-vs-rand data performs poorly on this task, indicating that the generated responses are sufficiently relevant to the context to yield a higher score than a random response. This is consistent with the evaluation results reported by Zhang et al. (2019b), which shows that DialoGPT receives higher relevancy score in a human evaluation. However, the feedback prediction models, Width, Depth and Updown, show much higher accuracy in the human-vs-generated task, even though they were not trained on any generated responses. This implies that the ranking models predict that DialoGPT's generated responses may not be as proactive or as engaging as human responses. Finally, the model trained with both random and generated responses perform well on both human-vs.-fake tasks, but not well on the human-vs.-human feedback ranking tasks. This indicates that the models are complementary to each other, motivating us to build an ensemble model.

4.3 Ensembling Models

Reddit test data. The feedback and the human-like models are combined following Eq. 7 and evaluated using different test sets, as shown in Table 6. For testing on feedback K, where K is Width, Depth or Updown, we set $w_i = 1$ if $i = K$ and 0 otherwise. For human vs. fake, we set $w_K = 1/3$ for all three feedback models. Although the ensemble model's accuracy is not the highest for any of the test sets, it performs reasonably well on all of them.

Human overall preference. We also test the correlation between the ensemble model and human overall preference, using the human annotations introduced in Section 3.2. As shown in Table 8, adding the human-like model π_0 improves the model performance, indicated by the comparison between the model $\pi_0 \sum_K w_K s_K$ and $\sum_K w_K s_K$. Among the three feedback modes, human preference correlates best with Updown. Presumably, Upvotes (or "Likes"), is more directly tied to human preference than Width or Depth. However, the other two metrics are useful as well.The fitted coefficients of the $\sum_K w_K s_K$ model implies the overall preference is a combination of these modes, favoring replies that can prolong a dialog session ($w_{\text{Depth}} = 0.48$), that are likely to be upvoted ($w_{\text{Updown}} = 1.0$) and that do not target too

Model	Trained on	Tested on				
		Human vs. Human			Human vs. Fake	
		Width	Depth	Updown	Rand	Generated
Human feedback	Width	0.764	0.693	0.601	0.517	0.644
	Depth	0.749	0.701	0.588	0.512	0.647
	Updown	0.659	0.602	0.683	0.526	0.667
Human-like	Rand	0.558	0.552	0.522	0.843	0.413
	+ Generated	0.560	0.558	0.522	0.864	0.880
Ensemble	-	0.746	0.675	0.666	0.758	0.821

Table 6: Pairwise accuracy of DIALOGRPT models. Darker cell color indicates better performance.

Dataset	Method	Hits@k
Reddit (k >5,n=k)	BLEU1	0.651
	BERTScore	0.685
	BLEURT	0.714
	BM25	0.309
	ConvRT	0.760
	Dialog ppl.	0.560
	Reverse dialog ppl.	0.775
	DIALOGRPT	**0.886**
DailyDialog (k=1,n=19)	BM25	0.182
	ConvRT	0.380
	Dialog ppl.	0.176
	Reverse dialog ppl.	0.457
	DIALOGRPT	**0.621**
Twitter (k=1,n=19)	BM25	0.178
	ConvRT	0.439
	Dialog ppl.	0.107
	Reverse dialog ppl.	0.440
	DIALOGRPT	**0.548**
PersonaChat (k=1,n=19)	BM25	0.117
	ConvRT	0.197
	IR Baseline	0.213
	Starspace	0.318
	KV profile memory	0.349
	Dialog ppl.	0.108
	Reverse dialog ppl.	0.449
	DIALOGRPT	**0.479**

Table 7: Performance ranking k gold and n distractor responses. DIALOGRPT is trained on Reddit human-vs-rand dataset, and is *zero-shot* for other datasets in the table.

	Acc.	ρ
Dialog ppl.	0.539 (0.033)	0.082 (0.060)
Reverse dialog ppl.	0.548 (0.031)	0.094 (0.056)
DIALOGRPT		
$\pi_0 s_{\text{Width}}$	0.749 (0.008)	0.465 (0.012)
$\pi_0 s_{\text{Depth}}$	0.762 (0.009)	0.467 (0.013)
$\pi_0 s_{\text{Updown}}$	0.760 (0.008)	0.470 (0.013)
$\sum_K w_K s_K$	0.629 (0.014)	0.201 (0.019)
$\pi_0 \sum_K w_K s_K$	**0.792** (0.010)	**0.518** (0.015)

Table 8: Performance of human overall preference prediction measured by acurracy (Acc.) and Pearson correlation (ρ). Values are reported in form "average (standard error)" of 10-fold cross validation results.

broad an audience ($w_{\text{Width}} = -0.50$).

Improving generation model. Even when the generative model (i.e. DialoGPT) is held constant, DIALOGRPT improves candidate ranking in comparison to perplexity-based methods. This indicates that incorporating human feedback information into response generation ranking methods can yield improvements over methods that rely solely on measures of relevancy.

5 Related Work

Dialog hypothesis ranking. Earlier work has explored the use of generation probability $P(h|x)$ or perplexity for hypothesis ranking. Li et al. (2016) combine this with reverse dialog probability to consider mutual information (MMI) in ranking dialog response hypotheses Gao et al. (2019b) adds style intensity for stylized response generation. Another line of works (Henderson et al., 2019a; Humeau et al., 2019) encodes context and candidate as vectors and use their similarity for ranking. Some systems (Zhou et al., 2018; Gao et al., 2020) employ a set of features to rank hypotheses, e.g., local cohesion, global coherence, empathy matching, and retrieval matching.

Reference-based quality measure is also used to estimate the quality of response, although this is not applicable on-the-fly. BLEU (Papineni et al., 2002) is a classic metric measuring the sentence similarity using ngram overlap. BERTScore (Zhang et al., 2019a) uses BERT contextualized word embeddings, instead of ngrams. BLEURT (Sellam et al., 2020) directly measures sentence-level similarity, initialized with BERT and then trained on millions of synthetic examples.

Contrastive Learning focuses on the relation between samples or labels. Hadsell et al. (2006) learns representations using a contrastive loss function which pulls neighbors together and pushes apart non-neighbors in the learned space. Gao et al. (2019a) designed a loss function to reduce the distance between matched context and response in contrast to the random pairs. Chen et al. (2020)

proposed a contrastive learning framework, establishing a new state-of-the-art for image classification.

Social sciences and social-media NLP: Glenski and Weninger (2017) model each user separately and predict their interaction for a given post using features including existing upvotes/downvotes, rank, and bag of words. Stoddard (2015) models upvotes as a time-series function of content quality, displaying position, age and score of the post and shows that popularity is positively correlated with quality, though articles of similar quality can have very different numbers of upvotes. Lakkaraju et al. (2013) studied resubmissions to decompose article popularity into the quality of the content and the appeal of the title. They find that textual features of the title significantly affect popularity.

6 Conclusion

We leverage Reddit human feedback data to build and release a large-scale training dataset for feedback prediction.We trained GPT-2 based models on 133M pairs of human feedback data and demonstrate that these models outperform several standard baselines. In particular, the conventional dialog perplexity baseline shows little predictive power on Reddit human feedback data. We ensemble the feedback prediction models and a human-like scoring model to rank the machine generated dialog responses. Human evaluation shows that human preference is improved with our ranking method. For the future work, we suggest to integrate the ranking models and generation model, e.g., in beam search stage or reinforcement learning using ranking score as reward signal.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Michel Galley, Chris Brockett, Xiang Gao, Bill Dolan, and Jianfeng Gao. 2018. End-to-end conversation modeling: Dstc7 task 2 description. In *DSTC7 workshop (forthcoming)*.

Xiang Gao, Michel Galley, and Bill Dolan. 2020. Mixingboard: a knowledgeable stylized integrated text generation platform. *arXiv preprint arXiv:2005.08365*.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019a. Jointly optimizing diversity and relevance in neural response generation. *NAACL-HLT 2019*.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019b. Structuring latent spaces for stylized response generation. In *Proc. of EMNLP*, pages 1814–1823.

Eric Gilbert. 2013. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 803–808.

Maria Glenski and Tim Weninger. 2017. Predicting user-interactions on reddit. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 609–612.

H Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulic, and Tsung-Hsien Wen. 2019a. A repository of conversational datasets. In *Proceedings of the Workshop on NLP for Conversational AI*. Data available at github.com/PolyAI-LDN/conversational-datasets.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Ivan Vulić, et al. 2019b. Convert: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. 2013. What's in a name? understanding the interplay between titles, content, and communities in social media. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Chunyuan Li, Xiang Gao, Yuan Li, Xiujun Li, Baolin Peng, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. *arXiv preprint arXiv:2004.04092*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL*, pages 110–119.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856.

Matthew J Salganik and Duncan J Watts. 2008. Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social psychology quarterly*, 71(4):338–355.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *Proc. of ACL*.

E. Isaac Sparling and Shilad Sen. 2011. Rating: How difficult is it? In *Proceedings of the Fifth ACM Conference on Recommender Systems*, page 149–156, New York, NY, USA. Association for Computing Machinery.

Greg Stoddard. 2015. Popularity and quality in social news aggregators: A study of reddit and hacker news. In *Proceedings of the 24th international conference on world wide web*, pages 815–818.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *Proc. of ICLR*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pages 1813–1823.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *Proc. of ACL*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 654–664.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The design and implementation of xiaoice, an empathetic social chatbot. *arXiv preprint arXiv:1812.08989*.