

Balancing Multi-Domain Corpora Learning for Open-Domain Response Generation

Yujie Xing¹, Jinglun Cai^{2*}, Nils Barlaug¹, Peng Liu¹, Jon Atle Gulla¹

¹Department of Computer Science, Norwegian University of Science and Technology

²Amazon AWS AI

{yujie.xing, nils.barlaug, peng.liu, jon.atle.gulla}@ntnu.no
cjinglun@amazon.com

Abstract

Open-domain conversational systems are assumed to generate equally good responses on multiple domains. Previous work achieved good performance on the single corpus, but training and evaluating on multiple corpora from different domains are less studied. This paper explores methods of generating relevant responses for each of multiple multi-domain corpora. We first examine interleaved learning which intermingles multiple corpora as the baseline. We then investigate two multi-domain learning methods, labeled learning and multi-task labeled learning, which encode each corpus through a unique corpus embedding. Furthermore, we propose Domain-specific Frequency (DF), a novel word-level importance weight that measures the relative importance of a word for a specific corpus compared to other corpora. Based on DF, we propose weighted learning, a method that integrates DF to the loss function. We also adopt DF as a new evaluation metric. Extensive experiments show that our methods gain significant improvements on both automatic and human evaluation. We share our code and data for reproducibility.¹

1 Introduction

Recent work has achieved improvements in general performance for open-domain response generation (Vinyals and Le, 2015; Serban et al., 2017; Li et al., 2016; Xu et al., 2018). However, most studies are restricted to single-corpus training and evaluating, while there lacks studies for training and evaluating with multiple corpora from different domains. Single-corpus training has intrinsic limitations. For example, a corpus of everyday chats, e.g., the PersonaChat corpus (Dinan et al., 2019), does not cover technical topics discussed in

* This work was done prior to the author joining Amazon.

¹https://github.com/yujie-xing/Balancing_Multi_Domain_Corpus_Learning_for_Open_Domain_Response_Generation

Context	What are you going to do on the remote system exactly?
PersonaChat	I am going to be a pilot. I am going to fly planes.
4 corpora (concatenated)	I am going to go to the beach.

Table 1: Irrelevant responses generated from fine-tuned GPT-2. The GPT-2 model is fine-tuned respectively on PersonaChat / concatenated 4 corpora (OpenSubtitles, Twitter, Ubuntu, PersonaChat)

Fine-tune corpus	Test set			
	OSDB	Twitter	Ubuntu	PersonaChat
PersonaChat	478.8	159.6	264.7	19.6
4 corpora (concatenated)	392.8	110.7	199.2	19.0

Table 2: Imbalanced perplexity performance of fine-tuned GPT-2. The GPT-2 model is fine-tuned on PersonaChat / concatenated 4 corpora (OpenSubtitles, Twitter, Ubuntu, PersonaChat)

Ubuntu chatlogs (Lowe et al., 2015). A conversational system that learns only from PersonaChat or from multiple corpora without an appropriate technique is not likely to generate relevant responses for certain topics (see Table 1). Therefore, it is necessary for an open-domain conversational system to learn from multiple corpora, and to learn with good techniques.

Furthermore, the case of using a single small-scale open-domain corpus has apparent weaknesses. A common way of dealing with a small-scale corpus is through fine-tuning (Li et al., 2016; Akama et al., 2017; Chu et al., 2017). Fine-tuning on a single corpus tends to make the model overfit on that specific corpus while performing worse on other corpora. Table 2 shows the result of a GPT-2 model gaining good performance on PersonaChat while performing poorly on other corpora.

This paper explores how to train and evaluate on multiple corpora from different domains for the open-domain response generation task. We propose

several methods to make a model generate relevant responses for each of the multiple corpora.

Since simply training multiple corpora one by one does not solve the imbalanced performance (as shown in Table 1 and 2), we first investigate *interleaved learning*, a method that intermingles the training data instead of simply concatenating, which ensures a model learns from all corpora evenly. We use this method as a baseline. Additionally, we explore two multi-domain learning methods: *labeled learning* and *multi-task labeled learning*. Labeled learning comes from a control technique in response generation (Li et al., 2016; Johnson et al., 2017; Yang et al., 2017). Previous works focus on controlling persona and style, while our method controls corpus’s information with the corpus embedding. Multi-task labeled learning is inspired by works of domain adaption (Luan et al., 2017; Niu and Bansal, 2018; Chu and Wang, 2018), where multiple losses from both the corpus classifier and response generator are minimized. To the best of our knowledge, this paper is the first that uses corpus embeddings on the open-domain response generation task for multiple corpora.

Furthermore, we propose a novel *weighted learning* with Domain-specific Frequency (DF). DF is a word-level importance weight (Leopold and Kindermann, 2002) that assigns different weights (importance) to the same words from different corpora. In the training process, we weight the loss of a model with DF, so that the model focuses on the most important words for a specific corpus.

For automatic evaluation metrics, we eliminate the stop words and use ROUGE-1 (precision, recall, F1) (Lin, 2004) to measure the **relevance** of the generated responses. In addition, we adopt DF to see how relevant the generated response of a model is to a specific corpus. We will explain DF as an evaluation metric in Section 4.4. Results show that for overall performance, the best method (weighted learning) improves 27.4% on precision, 45.5% on recall, and 34.1% on F1. Further, it has at least 20.0% higher DF, stating that it uses more important words from the “correct” corpus. We also conduct an extensive human evaluation on 2400 generated responses. The human evaluation shows a highly significant ($p < 0.001$) improvement on all of our proposed methods, especially the weighted learning method.

We summarize our work as follows:

- We explore the problem of training and eval-

uating on multiple corpora from different domains for open-domain response generation. The task is to make the conversational models generate relevant responses for **each** corpus.

- We examine several multi-domain corpora learning methods for their ability to solve the proposed task.
- We propose Domain-specific Frequency (DF) as in weighted learning and as an evaluation metric. DF distinguishes important words for each corpus and helps a model to focus on these important words in the training process.

2 Related Work

Open-Domain Response Generation Recent work of open-domain response generation generally follows the work of Ritter et al. (2011) where the task is treated as a machine translation task, and many of them use a Seq2Seq structure (Sutskever et al., 2014) following previous work (Vinyals and Le, 2015; Shang et al., 2015; Sordoni et al., 2015). In recent years, substantial improvements have been made (Serban et al., 2017; Li et al., 2016; Wolf et al., 2019), and embeddings are used to control response generation on extra information such as persona (Li et al., 2016), profiles (Yang et al., 2017), coherence (Xu et al., 2018), emotions (Huang et al., 2018), and dialogue attributes like response-relatedness (See et al., 2019). However, there is a lack of work that uses embeddings to control response generation over multiple corpora. Our work follows the common models of open-domain conversational systems, while we study the problem of multiple corpora of different domains.

Multi-Domain Learning and Domain Adaption Multi-domain learning aims at making a conversational model learn from multiple domains to prevent the performance from degrading due to domain differences (Ben-David et al., 2007). There are two categories of solutions for multi-domain learning (Joshi et al., 2012): (i) capturing domain-specific characteristics in the parameters (Daumé III, 2007); (ii) capturing the relationship among different domains (Saha et al., 2011).

Some work of natural language generation and machine translation is related to multi-domain learning. Luan et al. (2017) and Niu and Bansal (2018) use multi-task learning for domain adaption respectively on speaker-role and politeness. Wen

et al. (2016) and Akama et al. (2017) utilizes fine-tuning as a common way of domain adaption for language generator and style transferer. For machine translation, in order to deal with the mixed-domain parallel corpus, Zeng et al. (2018) adjust the weights of target words in the training objective based on their relevance to different domains. We differ in that we propose DF and we deal with the response generation task. Chu et al. (2017) propose mixed fine-tuning, which adds the out-of-domain pre-training data to the fine-tuning dataset, and they observe an improvement of performance. In this paper, we also mix small-scale fine-tuning datasets with out-of-domain training data, while the data we add is not necessarily used during pre-training. Shi et al. (2015) state that fine-tuning can be done by placing the corpus to be fine-tuned at the end of the entire corpus, which is an extension of curriculum learning proposed by Bengio et al. (2009). We also explore how the order of multiple corpora influences the result, but our focus is on balancing performance. Recently, Smith et al. (2020) investigated blending conversational skills with knowledge and empathy skills, where they mix 3 corpora. They focus on selecting appropriate skills and they propose a blended corpus with labels, while we focus on generating responses that are most relevant to a specific corpus.

3 Base Models

We use two base models: an LSTM Seq2Seq model with attention (Hochreiter and Schmidhuber, 1997; Sutskever et al., 2014; Bahdanau et al., 2015) and a pre-trained GPT-2 model (Radford et al., 2019). The LSTM Seq2Seq model with attention is a common model for conversational systems (Li et al., 2016; See et al., 2019), and the GPT2 model is a state-of-the-art model for the response generation task (Zhang et al., 2020; Zhao et al., 2020).

The basic task of response generation is to predict the next word given the past and current words of the context and response, and to make the generated response as similar to the original response as possible. The task can be described as follows. Probability of response Y given context X is predicted as:

$$P(Y|X) = \prod_{t=1}^n P(y_t|y_1, \dots, y_{t-1}, X), \quad (1)$$

where $X = x_1, \dots, x_m$ and $Y = y_1, \dots, y_n$ is a context-response pair.

3.1 LSTM Seq2Seq Model with Attention

We simplify an LSTM with attention unit as $LSTM^*$ since it is well introduced in previous work (Li et al., 2016). We calculate the hidden vector h_t at step t as:

$$h_t = LSTM^*(h_{t-1}, E(z_t)), \quad (2)$$

where $h_{t-1} \in \mathbb{R}^{dim}$ is the hidden vector at step $t - 1$, dim is the dimension of hidden vectors, and $E(z_t)$ is the word embedding for word $z_t \in (x_1, \dots, x_m, y_1, \dots, y_{n-1})$. We apply dot multiple in the attention mechanism when calculating the context vector c_t :

$$c_t = H \cdot (\text{softmax}(H^\top \cdot h_t))$$

where $H \in \mathbb{R}^{d \times m}$ is the concatenation of hidden vectors from the encoder. c_t is input to the next step $t + 1$ in the decoder. Each token's hidden vector h_t in the decoder is combined with c_t through a linear layer and an activation to predict the next token.

3.2 GPT-2

As for GPT-2, we follow the adaption of Wolf et al. (2019). The transformer block of GPT-2 captures the relation of multiple words in one sentence, which largely follows Vaswani et al. (2017). The hidden vector to be input to the transformer block is calculated as:

$$h_{0[t]} = E(X, Y_{[1:t]}) + (E_0, E_1) + W_p, \quad (3)$$

where $Y_{[1:t]}$ is (y_1, \dots, y_t) , $E(X, Y_{[1:t]})$ is the sub-word embedding for context X and response $Y_{[1:t]}$. E_0 and E_1 are dialogue-state embeddings, which tutor the model to distinguish between contexts and responses. W_p is a pre-trained position embedding. The probability of the subword to generate is then calculated as:

$$h_{[t]} = \text{transformer_block}(h_{0[t]}) \quad (4)$$

$$P(y)_{t+1} = \text{softmax}(E^\top(h_{[t]})), \quad (5)$$

where $y \in V$, and V stands for the sub-word vocabulary. We simplify the structure of transformer block as *transformer_block*. In the block, a mask is filled in the attention matrix, which bans past words from attending to future words. This ensures that the model follows the traditional language modeling. The hidden vector of t_{th} sub-word is used to generate the probability distribution for the vocabulary ($P(y)$, $y \in V$) for $(t + 1)_{th}$ sub-word. E^\top means that the model uses the sub-word embeddings in calculating sub-word probabilities for generation (Press and Wolf, 2017).

4 Proposed Methods

4.1 Interleaved Learning

Interleaving is a concept in cognitive psychology proven to be efficient for learning (Kornell and Bjork, 2008): intermingling learning material of different topics helps students to gain better learning results than learning the material topic by topic. Previous work from machine learning also shows that training order greatly influences the performance (Bengio et al., 2009). When the training is conducted on a simple concatenation of multiple corpora, the model tends to concentrate on the last corpus (Shi et al., 2015). To address this issue, we propose interleaved learning as an alternative: each time we collect one context-response pair from each of the corpora, and we randomly shuffle them. For example, if there are 3 corpora $(a_1, a_2, \dots), (b_1, b_2, \dots), (c_1, c_2, \dots)$ where a_i, b_i and c_i are context-response pairs, the resulting mixed corpus might be $(b_1, a_1, c_1, c_2, b_2, a_2, \dots)$. Interleaved learning guarantees that the combined corpus is evenly distributed, which helps the model learn from multiple corpora evenly.

4.2 Labeled Learning

We propose our labeled learning as follows: each corpus is assigned a randomly initialized unique embedding, and the conversational model learns these embeddings together with conversations during the training period. We denote these embeddings as “corpus embedding”, or E_c . A model captures each corpus’s characteristics through the corpus embedding and uses it to control the generated responses. To know which corpus embedding to use, each context is labeled with which corpus it comes from, and these labels are provided to the model both in the training and generation period. We propose an approach for each of our base models for encoding corpus embeddings.

For the LSTM model, following Li et al. (2016), we input the corpus embedding E_c into the first layer of the decoder LSTM at every step, together with the response words. Calculation of a hidden vector h_t in the decoder LSTM is then adapted to:

$$h_t = LSTM^*(h_{t-1}, E(y_t), E_c). \quad (6)$$

The structure is illustrated in the dashed red rectangle of Figure 1a.

For the GPT-2 model, our method is based on Wolf et al. (2019). Instead of two kinds of dialogue-

state embeddings (context embedding E_0 and response embedding E_1), we replace the response embedding with corpus embeddings E_c . As a result, the model is aware of which corpus the response belongs. Calculation of a hidden vector to be input to the transformer block is adapted to:

$$h_{0[t]} = E(X, Y_{[1:t]}) + (E_0, E_c) + W_p. \quad (7)$$

The structure is illustrated in Figure 1b.

4.3 Multi-Task Labeled Learning

Labeled learning needs corpus labels for both training and generation processes. To avoid providing labels in the generation process, we combine multi-task learning with labeled learning on multiple corpora. Here, the conversational model has to predict by itself which corpus a context belongs to, which is expected to result in worse performance, but less information is required. In the encoder, we have a classifier layer that uses the sum of hidden vectors from the encoder ($\sum H$) to predict the corpus of a context. The loss of the classifier is calculated as:

$$\mathcal{L}_c = -\log \left(\text{softmax} \left(\left(\sum H \right) \cdot W_{[c]} \right) \right), \quad (8)$$

where $W_{[c]} \in \mathbb{R}^{dim}$ is the part from the classifier layer for target corpus c . \mathcal{L}_c is summed up with the loss from the response generator. The predicted corpus embedding is input into the decoder like labeled learning (see Section 4.2). The simplified structure is illustrated in Figure 1a.

4.4 Document-specific Frequency (DF)

We propose Domain-specific Frequency (DF) to measure how important a word is with respect to a different corpus under a collection of corpora. DF is used for weighted learning and evaluation. It is calculated as follows:

$$f(w)_d = \text{freq}(w)_d - \min_v \{ \text{freq}(v)_d \} \quad (9)$$

$$df(w)_d = \begin{cases} 0 & f(w)_d = 0 \\ \frac{f(w)_d}{\sum_{d \in D} f(w)_d} & f(w)_d \neq 0 \end{cases} \quad (10)$$

$$DF(w)_d = \frac{df(w)_d}{\max_v \{ df(v)_d \}}, \quad (11)$$

where $\text{freq}(w)_d$ is the relative frequency of a word w in a corpus d , and D represents the set of all corpora. It is easy to see from Equation 10 that $DF(w)_d$ represents the importance of word w for corpus d compared to other corpora. For a word w that frequently appears in corpus d but seldom

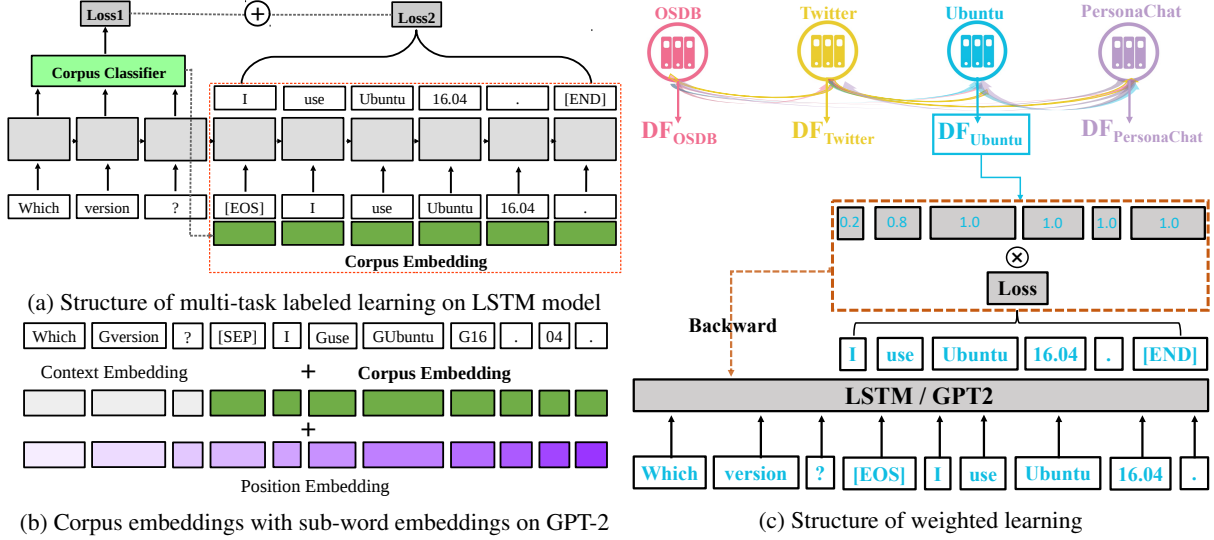


Figure 1: Adapted models with labeled learning, multi-task labeled learning and weighted learning

in other corpora (e.g., “upgrade” from Ubuntu corpus), $\sum_{d \in D} f(w)_d$ is close to $f(w)_d$, making $DF(w)_d$ approach 1. A word that frequently appears in all corpora (e.g., “I”, “you”) is punished, resulting in a lower $DF(w)_d$. A word that seldom appears in corpus d but frequently appears in other corpora (e.g., “music” seldom appears in Ubuntu corpus, but is common in other corpora) has the lowest $DF(w)_d$. Words that appear minimal times (e.g., once) in a corpus are ignored with Equation 9. Words that appear few times (e.g., twice or three times) are not dealt with, yet they are not of great influence in our experiments. We apply a normalization in the final step (Equation 11) to make $DF(w)_d$ of each corpus d range from 0 to 1.

We show $DF(w)_{\text{Ubuntu}}$ and $DF(w)_{\text{PersonaChat}}$ of some words in Table 3. We also show the results of TF-IDF (log normalization variant), a commonly used word importance weight, as a comparison. As expected, for the corpus Ubuntu and PersonaChat, most unique words w have very different $DF(w)_{\text{Ubuntu}}$ and $DF(w)_{\text{PersonaChat}}$. Unique words of each corpus get the highest values for the corresponding corpus, like “upgrade” for the Ubuntu corpus and “music” for the PersonaChat corpus; these words receive the lowest values for *incorrect* corpora, like “upgrade” for PersonaChat and “music” for Ubuntu. The stress on unique words makes DF more suitable for our task.

Weighted Learning with DF Weighted learning weights the loss of the predication y' for each target word w using $DF(w)_d$. In the training period, each context is labeled with the corpus d it belongs

Word	TF-IDF(%)		DF(%)		$\alpha DF_{(\alpha=100)}$	
	Ubuntu	PersonaChat	Ubuntu	PersonaChat	Ubuntu	PersonaChat
i	100.0	62.6	20.8	42.1	2.6	7.3
to	64.6	32.8	26.9	24.9	3.8	3.1
it	83.2	21.7	38.5	14.5	5.1	2.1
laptop	5.4	0.2	89.8	4.5	76.0	1.0
upgrade	6.8	0.1	95.6	0.4	91.2	1.0
file	15.7	0.1	96.0	0.3	86.4	0
windows	12.2	0.1	97.1	0.1	86.3	1.0
ubuntu	27.5	0	99.9	0	99.5	0
teacher	0.1	2.2	0.7	77.8	1.0	53.5
music	1.5	7.6	4.8	82.9	1.2	49.1
travel	0.1	3.1	0.3	88.9	1.0	57.1
hobby	0.1	1.6	0.6	94.3	1.1	81.7
hiking	0	1.5	0	97.6	0	91.8

Table 3: Normalized TF-IDF (%), DF (%) and αDF of some words for Ubuntu and PersonaChat (more examples on other corpora can be found in Section A)

to, so that the model can use the $DF(w)_d$ of the corresponding corpus. Here DF is calculated only on the training sets. In the generation step, corpus labels are not provided, so DF is not used. The loss is weighted as follows :

$$\mathcal{L}_{\text{weighted}} = DF(w)_d \cdot (-\log(\text{softmax}(y'_w))), \quad (12)$$

where y'_w represents the model’s predicted score for the target word w . With the weighted loss, the model concentrates on words that are important to the corpus of the current context, and focuses less on frequent words or words that are not important to the current corpus. The structure is illustrated in Figure 1c.

Evaluation with DF For the generated responses to be relevant to a specific corpus, they have to

be similar to that corpus, which includes using important words of that corpus (e.g., responses generated for the Ubuntu corpus should have more technical words than other corpora). Thus, we propose DF as an evaluation metric that shows to what extent the generated responses use important words of the corresponding corpus. We want to decrease the influence of common words like “i”, “to”, etc., and thus address the important words. So we adopt exponential DF with α as the base (α DF):

$$\alpha\text{DF}(w)_d = \begin{cases} 0 & \text{DF}(w)_d = 0 \\ \alpha^{\text{DF}(w)_d} & \text{DF}(w)_d \neq 0, \end{cases} \quad (13)$$

where α is a constant. $\alpha\text{DF}(w)_d$ rescales $\text{DF}(w)_d$ by exponent with α as a base. In our experiments, we set α to be 100, which transforms the range of the metric from $(0, 1)$ to $(0, 100)$. This makes the difference between high and low αDF more significant than DF and gives a 100-scale score. For each corpus $d \in D$, we average $\alpha\text{DF}(w)_d$ on word w from the generated responses of each test set, which gives us αDF_d scores ($d \in D$) for each test set. Ideally, the generated responses of a specific corpus d should have a higher αDF_d score and lower $\alpha\text{DF}_{\bar{d}}$ score ($\bar{d} \in \{d' \in D \mid d' \neq d\}$). For example, generated responses of the Ubuntu test set should have a higher $\alpha\text{DF}_{\text{Ubuntu}}$ score, while a lower $\alpha\text{DF}_{\overline{\text{Ubuntu}}}$ score ($\overline{\text{Ubuntu}} \in \{d' \in D \mid d' \neq \text{Ubuntu}\}$). αDF_d scores for responses from the original test sets are the standard scores.

We show $\alpha\text{DF}(w)_{\text{Ubuntu}}$ and $\alpha\text{DF}(w)_{\text{PersonaChat}}$ (calculated purely on test set) in Table 3. As expected, αDF has a more significant difference between important words and common words.

Is DF a Legal Evaluation Metric? Although DF is used for both weighted learning and evaluation, we see DF as a suitable evaluation metric for our task and not biased in favor of weighted learning due to: 1) A word receives multiple DF values in the training process given the corpus that a context belongs to; 2) in the generation process, DF is never used. 3) In the evaluation process, DF can be calculated purely on the test sets. Note that since a word receives multiple DF values in the training step, it is equivalently likely for the model trained with weighted learning to be influenced by DF weights of **incorrect** corpus. Above all, in the evaluation step, if the trained model is influenced more by DF weights from the correct corpus, it already means that the model is good

at distinguishing which corpus a given context is from, thus is suitable for our task.

5 Experiment Setup

5.1 Datasets

Data Collection We collected 4 commonly used English corpora of different domains from the ParlAI platform (Miller et al., 2017): OpenSubtitles corpus (OSDB)² (Lison et al., 2018), Twitter corpus³ (Miller et al., 2017), Ubuntu chatlogs corpus (Lowe et al., 2015)⁴, and PersonaChat corpus (Zhang et al., 2018) from the NeurIPS 2018 ConvAI2 Challenge (Dinan et al., 2019). Each corpus contains 250K context-response pairs, as much as the size of the original PersonaChat used in ConvAI2 competition. This gives us 1M context-response pairs in total. The corpus for training is a combination of these 4 corpora. For comparison, we have a single corpus–PersonaChat–trained on both base models. For testing, each of the 4 corpora has a test set of 30K context-response pairs, which is the same size of the test set of PersonaChat.

The OpenSubtitles corpus (OSDB) is a noisy dataset of film subtitles. We removed films that belonged to genres that usually had few conversations, such as musical and documentary films. We regarded two neighboring sentences as a context-response pair following Vinyals and Le (2015). The Twitter corpus contains one-turn dialogues extracted from Twitter. The original author has already cleaned it, so we only removed special symbols such as hashtags, Emojis, and @. The Ubuntu corpus contains dialogues about solving technical problems of Ubuntu. The PersonaChat corpus contains dialogues between two workers acting as specific personas; we focused on the dialogue part and ignored the persona part. This corpus allows us to compare our base models with state-of-the-art performance. These 4 corpora have very different characteristics, confirmed by the imbalanced performance of GPT-2 fine-tuned on a single corpus (see Table 2).

5.2 Training and Decoding

We used Pytorch (Paszke et al., 2017) to implement the LSTM Seq2Seq model with attention and the pre-trained GPT-2 models. For GPT-2, we adapted

²<http://www.opensubtitles.org/>

³https://github.com/Marsan-Ma/chat_corpus/

⁴<https://github.com/rkadlec/ubuntu-ranking-dataset-creator>

Model	Corpus / Method	Test set														
		OSDB			Twitter			Ubuntu			PersonaChat			Overall		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
LSTM	PersonaChat (single)	11.8	8.9	8.1	12.4	8.6	8.9	12.1	8.1	7.7	56.7	43.4	45.8	23.2	17.2	17.6
	Concatenated	11.0	7.7	7.2	15.7	10.9	11.4	36.5	17.8	20.1	57.7	44.0	46.4	30.2	20.1	21.3
	Interleaved	24.1	10.1	11.7	24.3	12.5	14.9	58.4	24.9	29.6	56.1	41.5	44.3	40.7	22.3	25.1
	Labeled	23.9	10.1	11.3	24.5	13.2	15.5	61.6	26.5	31.6	56.4	43.0	45.4	41.6	23.2	26.0
	Multi-task Labeled	23.2	9.6	11.1	23.2	12.3	14.5	56.4	23.8	28.3	53.2	40.6	42.7	39.0	21.6	24.2
	Weighted	26.6	11.9	13.4	29.7	12.2	15.6	78.4	35.2	41.2	62.4	42.5	47.1	49.3	25.5	29.3
GPT-2	PersonaChat (single)	15.0	12.4	10.8	19.6	13.2	13.9	24.8	16.2	15.5	70.0	57.1	58.8	32.4	24.7	24.7
	Concatenated	17.4	14.1	12.6	24.5	16.4	17.2	35.0	22.5	22.4	66.8	55.4	56.3	35.9	27.1	27.1
	Interleaved	40.0	20.5	22.3	31.0	17.9	20.1	81.7	38.1	44.3	68.7	56.2	57.6	55.3	33.2	36.1
	Labeled	38.6	19.9	21.6	31.4	19.4	21.1	84.2	38.4	45.0	70.7	57.2	59.0	56.2	33.7	36.7
	Multi-task Labeled	38.4	19.8	21.4	31.2	18.6	20.6	80.9	37.8	43.8	68.0	56.0	57.3	54.6	33.0	35.8
	Weighted	41.9	21.2	23.4	39.9	18.4	22.3	86.8	43.3	48.6	69.0	53.2	55.8	59.4	34.0	37.5

Table 4: Precision, recall and F1 of ROUGE-1 (%) for baselines and proposed methods fine-tuned on 4 corpora (stop words eliminated)

Model	Corpus / Method	Test set													
		OSDB				Twitter				Ubuntu				PersonaChat	
		OSDB		PersonaChat		Twitter		PersonaChat		Ubuntu		PersonaChat		PersonaChat	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Test Set (Standard Score)		7.0	9.7	3.6	3.7	9.1	11.0	3.6	3.8	19.4	23.2	2.7	2.8	9.5	12.0
LSTM	PersonaChat (single)	2.9	3.4	9.2	9.9	2.8	3.4	8.6	9.2	2.7	3.1	8.6	9.1	11.9	12.6
	Concatenated	2.9	3.3	7.6	8.6	3.6	4.3	8.0	8.7	7.6	7.7	5.6	6.0	12.5	13.6
	Interleaved	3.9	4.1	5.0	5.3	4.7	4.9	4.1	4.5	11.8	11.3	3.7	4.0	11.5	12.5
	Labeled	3.9	4.2	5.0	5.3	5.0	5.3	3.9	4.3	12.5	11.8	3.4	3.8	12.1	13.1
	Multi-task Labeled	3.8	4.0	5.0	5.4	4.5	4.7	4.1	4.5	11.2	10.7	3.8	4.1	11.4	12.6
	Weighted	5.6	6.3	4.1	4.5	9.9	10.1	3.8	4.3	27.7	25.4	2.7	3.0	17.7	18.3
GPT-2	PersonaChat (single)	2.8	3.2	10.5	11.1	2.9	3.3	9.5	9.8	4.1	4.6	8.3	8.4	12.9	13.7
	Concatenated	3.1	3.6	8.8	9.4	3.3	3.9	8.2	8.7	6.5	7.1	7.0	7.4	12.1	13.0
	Interleaved	4.9	5.8	4.8	5.0	4.6	5.1	4.4	4.7	15.7	16.0	3.1	3.4	12.1	12.9
	Labeled	4.9	5.8	4.8	5.0	4.7	5.2	4.1	4.3	16.7	17.0	2.9	3.2	12.4	13.1
	Multi-task Labeled	4.8	5.7	4.8	5.1	4.6	5.1	4.4	4.6	15.5	15.8	3.1	3.4	12.1	12.9
	Weighted	6.0	7.5	4.1	4.4	8.1	8.8	3.7	4.1	25.7	24.4	2.4	2.6	16.0	17.1

Table 5: αDF_d scores for generated responses from multiple corpora. The columns “train” indicate train-set- αDF_d . The columns “test” indicate test-set- αDF_d .

our model from the implementation of the HuggingFace team⁵. The LSTM model has 4 layers and the dimension is 512. The training procedure was with a batch size of 256, learning rate of 1.0, dropout rate of 0.2, and gradient clip threshold of 5. The vocabulary size is 50000. GPT-2 has 12 layers, 12 heads, and the dimension is 768, the same as the pre-trained model. The training procedure was with Adam and we adopted a similar setup as Wolf et al. (2019): the batch size was 32, learning rate was 6×10^{-5} , $\beta_1 = 0.9$, $\beta_2 = 0.999$, L2 weight decay set to 0.01, learning rate linearly decreased to zero at the end. We followed these hyper-parameters to ensure state-of-the-art performance for the base models. We use the same hyper-parameters for both base models and models with our proposed methods, so the proposed methods work slightly

(but not much) worse than it should be. This is to avoid the extra improvement caused by hyper-parameters. We pre-trained the LSTM model on 3 large-scale corpora (OSDB, Twitter and Ubuntu) with interleaved learning until converging. GPT-2 is already pre-trained, so we directly used it for fine-tuning (details about pre-training convergence can be found in Section B). For decoding, we adopted greedy decoding for all the models to ensure an equal condition.

5.3 Evaluation

For automatic metrics, to measure the **relevance** of the generated responses, we eliminated punctuation and stop words, and adopted Rouge-1⁶ (precision, recall, F1) as multi-grams become meaningless

⁶We used implementation from <https://github.com/google-research/google-research/tree/master/rouge>.

⁵<https://huggingface.co/>.

without stop words. However, Rouge-1 compares the generated responses with the golden ones, while there is never a standard response for any context, so in addition to Rouge, we use α DF score that shows to what extent the generated responses use important words of the corresponding corpus, as stated in Section 4.4. Due to the limitation of automatic evaluation methods (Liu et al., 2016), we also conduct an extensive human evaluation on the relevance of generated responses to contexts (see Section 6.1 for details).

6 Results

Our base models achieve perplexity scores of 28.9 (LSTM model) and 19.6 (GPT-2) on the test set of the PersonaChat dataset from the ConvAI2 competition when fine-tuned with the single PersonaChat corpus (more details can be found in Section C). These results would likely advance the models to the second round in the competition.

Table 4 shows that models trained with our proposed methods gain better performance on Rouge than baselines. Baselines concentrate on the last trained corpus (PersonaChat), while with the proposed methods, performance is more balanced on multiple corpora. Weighted learning has the best overall performance on all metrics, and it performs especially well on the Ubuntu corpus, indicating that it might be good at distinguishing the unique technical words from the Ubuntu corpus. Labeled learning is the second best with stable improvement from interleaved learning, indicating that the corpus embeddings function as expected. Multi-task labeled learning has slightly worse performance than interleaved learning, indicating that predicting the corpus of a contexts is not easy, and wrong predictions result in worse performance.

Table 5 shows α DF_d scores for generated responses of each corpus. Full results can be found in Section E. We use both α DF_d calculated purely on the train set (train-set- α DF) and α DF_d calculated purely on the test set (test-set- α DF). The black scores are scores for the corresponding corpus (we expect high scores for these parts), while the grey scores are scores for non-related corpus—PersonaChat (we expect low scores for these parts). Note that scores for different corpora are in different scales. From the table, we can see that train-set-DF scores and test-set-DF scores are similar, and weighted learning always has the highest score, indicating that weighted learning distinguishes well

which corpus a context comes from. Labeled learning is the second best, indicating that the learned corpus embeddings help the model to use more important words of the corresponding corpus. Compared to the concatenated corpus, the improvement is at least 20%, while the decrease in PersonaChat is just 9% at most.

6.1 Human Evaluation

We conducted a human evaluation on all GPT-2 models: base models and models adapted with our proposed methods. We randomly picked 2400 responses: 400 different contexts evenly from 4 corpora with 6 responses generated by each of our models. 3 judges⁷ are asked to pick the most and the least relevant response(s) for the given context. The most relevant response(s) are given score 3, the least relevant response(s) are given score 1, and the other(s) are given score 2. Table 6 shows the overall scores of all GPT-2 based models. Table 7 shows the p-value for the t-test conducted between every two models. The overall scores of our proposed methods are all highly significantly ($p < 0.001$) higher than the concatenated models, especially the weighted learning method.

6.2 Response Examples

The generated responses from better methods are more relevant to the corresponding corpus, while worse methods cannot distinguish contexts from different corpora (e.g., they may answer any questions in a “PersonaChat” way). To show an intuition of the difference among our proposed methods, we present some response examples generated by GPT-2 in Section G.

6.3 Possible Limitations

Our proposed methods are meant to be able to work in most models, which is why we choose the most common conversational models as our base models. However, there are many variants of conversational models focusing on different aspects, such as integrating knowledge, avoiding dull responses, keeping the speech style, etc. We cannot ensure that our methods work for all of these variant models. Also, dialogues are always multi-turn, while we focus on a simpler task: single-turn response generation.

⁷Similar to previous work like Zhang et al. (2020), we have 3 judges. We have one random worker from <https://www.mturk.com/worker>, one bachelor student, and one graduate student. An example of the mTurk interface can be found in Section F.

Model \ Corpus	OSDB	Twitter	Ubuntu	PersonaChat	Overall
PersonaChat (single)	1.53	1.43	1.21	2.09	1.56
Concatenated	1.67	1.71	1.60	2.16	1.78
Interleaved	2.04	1.89	2.18	2.24	2.09
Labeled	2.10	2.10	2.32	2.24	2.19
Multi-task Labeled	2.05	1.98	2.11	2.24	2.10
Weighted	2.40	2.45	2.61	2.47	2.48

Table 6: Average scores of human evaluation for GPT-2 based models on each corpus

Model \ Model	PersonaChat	Concatenated	Interleaved	Labeled	Multi-Task Labeled	Weighted
PersonaChat	1.00	\	\	\	\	\
Concatenated	$2.54 \times 10^{-7**}$	1.00	\	\	\	\
Interleaved	$4.71 \times 10^{-34**}$	$2.09 \times 10^{-12**}$	1.00	\	\	\
Labeled	$1.08 \times 10^{-46**}$	$9.41 \times 10^{-21**}$	$1.18 \times 10^{-2*}$	1.00	\	\
Multi-task Labeled	$6.65 \times 10^{-35**}$	$6.96 \times 10^{-13**}$	8.86×10^{-1}	1.17×10	1.00	\
Weighted	$1.65 \times 10^{-103**}$	$2.86 \times 10^{-63**}$	$6.54 \times 10^{-26**}$	$1.59 \times 10^{-15**}$	$2.01 \times 10^{-25**}$	1.00

Table 7: P-value for t-test on overall human evaluation scores of GPT-2 based models, ** $p < 0.001$

Furthermore, the methods are trained and evaluated on English corpora. There can be a limitation on applying the methods to other languages.

7 Conclusions

We have experimented with 4 methods—interleaved learning (baseline), labeled learning, multi-task labeled learning, and weighted learning—to help common open-domain conversational systems generate relevant responses for multiple corpora of different domains. We adopted Rouge (precision, recall, F1) for auto evaluation. In addition, we used DF to evaluate how well a model uses relevant words for a corresponding corpus. We also did an extensive human evaluation. Our results show significant improvement in performance for our proposed methods, especially weighted learning. Future work of multi-turn response generation is potential. We have focused on one-turn response generation, while dialogue is naturally multi-turn so further research is needed.

Acknowledgements

This paper is funded by the collaborative project of DNB ASA and Norwegian University of Science and Technology (NTNU). We also received assist on computing resources from the IDUN cluster of NTNU (Själänder et al., 2019). We would like to thank Aria Rahmati, Zhirong Yang (Norwegian Research Council, 287284) and Özlem Özgöbek for their helpful comments.

References

- Reina Akama, Kazuaki Inada, Naoya Inoue, Sosuke Kobayashi, and Kentaro Inui. 2017. [Generating Stylistically Consistent Dialog Responses with Transfer Learning](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 408–412, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- 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*.
- Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. [Analysis of Representations for Domain Adaptation](#). In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. [Curriculum Learning](#). In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, pages 41–48. ACM.
- Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. [An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391. Association for Computational Linguistics.
- Chenhui Chu and Rui Wang. 2018. [A survey of domain adaptation for neural machine translation](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa

- Fe, New Mexico, USA. Association for Computational Linguistics.
- Hal Daumé III. 2007. [Frustratingly easy domain adaptation](#). In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.
- Emily Dinan, Varvara Logacheva, Valentin Lialykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. [The Second Conversational Intelligence Challenge \(ConvAI2\)](#). *arXiv:1902.00098 [cs]*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long Short-Term Memory](#). *Neural Comput.*, 9(8):1735–1780.
- Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. [Automatic Dialogue Generation with Expressed Emotions](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 49–54. Association for Computational Linguistics.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Mahesh Joshi, Mark Dredze, William W. Cohen, and Carolyn Rose. 2012. [Multi-Domain Learning: When Do Domains Matter?](#) In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1302–1312. Association for Computational Linguistics.
- Nate Kornell and Robert Bjork. 2008. [Learning concepts and categories is spacing the “Enemy of induction”?](#) *Psychological science*, 19:585–92.
- Edda Leopold and Jörg Kindermann. 2002. [Text categorization with support vector machines. how to represent texts in input space?](#) *Machine Learning*, 46(1):423–444.
- Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. [A Persona-Based Neural Conversation Model](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003. Association for Computational Linguistics.
- 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.
- Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. [OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. [How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132. Association for Computational Linguistics.
- Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. [The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems](#). In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294. Association for Computational Linguistics.
- Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 2017. [Multi-Task Learning for Speaker-Role Adaptation in Neural Conversation Models](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 605–614. Asian Federation of Natural Language Processing.
- Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. [ParLAI: A Dialog Research Software Platform](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84. Association for Computational Linguistics.
- Tong Niu and Mohit Bansal. 2018. [Polite Dialogue Generation Without Parallel Data](#). *Transactions of the Association for Computational Linguistics*, 6:373–389.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. [Automatic differentiation in PyTorch](#). In *NIPS-W*.
- Ofir Press and Lior Wolf. 2017. [Using the output embedding to improve language models](#). In *EACL*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Better Language Models and Their Implications](#). Technical report, Technical report, OpenAI.

- Alan Ritter, Colin Cherry, and William B. Dolan. 2011. [Data-Driven Response Generation in Social Media](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593. Association for Computational Linguistics.
- Avishek Saha, Piyush Rai, Hal Daumé III, and Suresh Venkatasubramanian. 2011. [Online Learning of Multiple Tasks and Their Relationships](#). In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 643–651. PMLR.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. [What makes a good conversation? How controllable attributes affect human judgments](#). *arXiv:1902.08654 [cs]*.
- Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. [A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues](#). In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. [Neural Responding Machine for Short-Text Conversation](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586. Association for Computational Linguistics.
- Yangyang Shi, Martha Larson, and Catholijn M. Jonker. 2015. [Recurrent Neural Network Language Model Adaptation with Curriculum Learning](#). *Comput. Speech Lang.*, 33(1):136–154.
- Magnus Sjölander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. 2019. [EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure](#).
- Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. [Can You Put it All Together: Evaluating Conversational Agents’ Ability to Blend Skills](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.
- Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. [A Neural Network Approach to Context-Sensitive Generation of Conversational Responses](#). In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205. Association for Computational Linguistics.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. [Sequence to Sequence Learning with Neural Networks](#). In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is All you Need](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Oriol Vinyals and Quoc Le. 2015. [A Neural Conversational Model](#). *arXiv:1506.05869 [cs]*.
- Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2016. [Multi-domain Neural Network Language Generation for Spoken Dialogue Systems](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 120–129. Association for Computational Linguistics.
- Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. [TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents](#). *arXiv:1901.08149 [cs]*.
- 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.
- Min Yang, Zhou Zhao, Wei Zhao, Xiaojun Chen, Jia Zhu, Lianqiang Zhou, and Zigang Cao. 2017. [Personalized Response Generation via Domain Adaptation](#). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’17*, pages 1021–1024. ACM.
- Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. 2018. [Multi-domain neural machine translation with word-level domain context discrimination](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 447–457, Brussels, Belgium. Association for Computational Linguistics.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing Dialogue Agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

Yizhe Zhang, Siqu 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.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. [Knowledge-grounded dialogue generation with pre-trained language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

A Comparison among TF-IDF, DF and α DF for 4 corpora on more example words

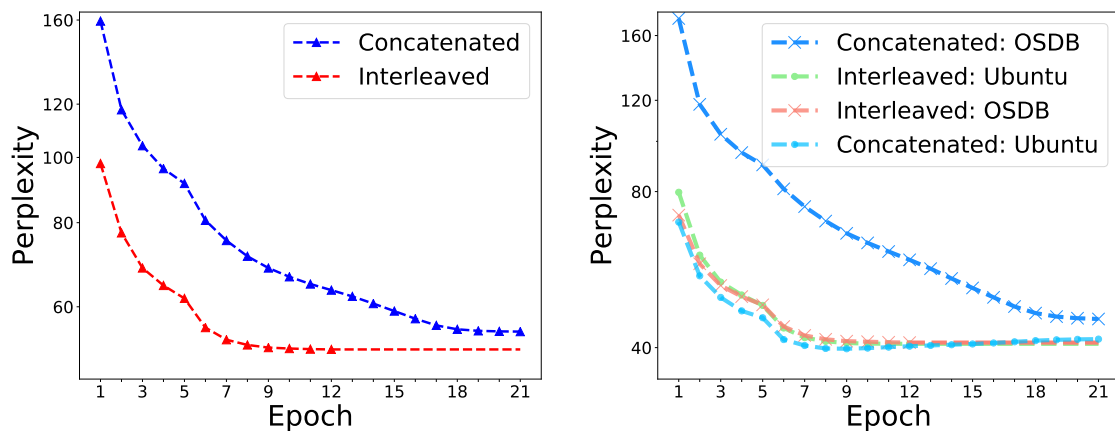
Word	TF-IDF(%)				DF(%)				α DF($\alpha=100$)			
	OSDB	Twitter	Ubuntu	PersonaChat	OSDB	Twitter	Ubuntu	PersonaChat	OSDB	Twitter	Ubuntu	PersonaChat
i	91.39	100.00	100.00	62.63	21.40	15.68	20.80	42.12	2.62	2.01	2.59	7.32
to	54.46	77.55	64.59	32.80	24.85	23.40	26.87	24.89	3.00	2.88	3.76	3.08
it	61.77	74.10	83.20	21.74	25.02	22.02	38.49	14.46	3.44	2.67	5.11	2.13
sword	0.64	0.17	0.01	0.08	68.37	13.74	0.26	17.63	63.29	1.37	1.00	1.15
forgive	2.41	0.48	0.16	0.06	75.35	14.37	5.44	4.84	50.96	1.58	1.19	1.05
hurry	5.21	0.52	0.09	0.08	88.39	6.67	1.48	3.45	63.53	1.32	1.15	1.04
darling	2.54	0.39	0.00	0.01	90.88	8.42	0.11	0.58	57.10	1.45	0	1.21
explain	1.27	0.00	0.00	0.11	91.33	0	0	8.67	94.14	0	0	1.06
tax	0.21	2.52	0.05	0.09	6.77	87.06	1.09	5.07	1.28	71.26	1.05	1.04
liberal	0.03	1.71	0.01	0.10	2.06	88.19	0.25	9.50	1.21	59.65	0	1.38
vote	0.41	6.08	0.10	0.11	6.07	90.68	0.78	2.47	1.12	80.22	1.02	1.09
trump	0.04	18.66	0.00	0.13	0.11	99.16	0.00	0.73	1.00	96.63	0	1.03
hillary	0.05	8.61	0.00	0.01	0.42	99.53	0	0.05	0	99.38	0	1.01
laptop	0.10	0.40	5.39	0.15	1.33	4.37	89.88	4.42	1.07	1.22	76.02	1.01
upgrade	0.03	0.47	6.85	0.03	0.24	3.75	95.63	0.37	1.01	1.06	91.24	1.03
file	0.64	0.55	15.65	0.05	2.29	1.44	96.02	0.26	1.11	1.04	86.36	0
windows	0.33	0.44	12.18	0.06	1.09	1.37	97.13	0.41	1.04	1.10	86.33	1.01
ubuntu	0.00	0.01	27.47	0.00	0	0.01	99.99	0	0	1.01	99.48	0
music	1.90	3.29	1.53	7.66	4.01	8.20	4.84	82.94	1.18	1.40	1.23	49.14
teacher	1.48	0.74	0.07	2.20	14.53	7.01	0.68	77.78	1.39	1.32	1.01	53.49
travel	0.42	0.91	0.05	3.07	3.91	6.89	0.28	88.92	1.27	1.36	1.01	57.15
hobby	0.10	0.27	0.04	1.56	1.94	3.03	0.57	94.46	1.13	1.00	1.09	81.71
hiking	0.03	0.09	0.00	1.52	0.85	1.45	0	97.70	0	1.09	0	91.76

Table 8: Normalized TF-IDF (%), DF (%) and α DF of more example words for 4 corpora

Example words are divided into five blocks. The first block has frequent words in all corpora, the second block has unique words from OSDB, the third block has unique words from Twitter, the fourth block has unique words from Ubuntu, and the fifth block has unique words from PersonaChat. The values of the corresponding corpus are marked with different colors.

From this table, it is clear that the commonly used word importance weight, TF-IDF, is not suitable for our task. This is due to the vast range of frequency, which leads to a relatively small penalty for IDF (Inversed Document Frequency) over words with too large TF (Term Frequency).

B Convergence time of pre-training LSTM model on large-scale corpora



(a) Overall perplexity of 3 corpora (OpenSubtitles, Twitter, Ubuntu) per epoch

(b) Perplexity of OSDB corpus and Ubuntu corpus per epoch

Figure 2: Convergence time of pre-training LSTM on large-scale corpora

In the pre-training period, it takes 21 epochs for the concatenated corpus to converge on the base LSTM model, while only 12 epochs with interleaved learning, which is 43% shorter. When trained on the concatenated corpus in the order of OSDB \rightarrow Twitter \rightarrow Ubuntu, it takes 20 epochs for the perplexity on OSDB and Ubuntu to be balanced, while with interleaved learning, it takes less than one epoch. For concatenated corpus, the performance of the Ubuntu corpus is sacrificed in order to balance the performance of the two corpora, which results in worse overall performance.

C Results of automatic evaluation with stop words

Model	Corpus / Method	Test set														
		OSDB			Twitter			Ubuntu			PersonaChat			Overall		
		Perp	BLEU	F1	Perp	BLEU	F1	Perp	BLEU	F1	Perp	BLEU	F1	Perp	BLEU	F1
LSTM	PersonaChat (single)	109.8	4.8	6.5	191.9	5.4	6.3	116.9	4.8	6.8	28.9	13.1	15.0	47.0	7.0	8.7
	Concatenated	57.0	4.8	6.3	111.4	5.9	6.1	50.0	5.1	6.8	27.8	13.2	15.1	36.8	7.2	8.6
	Interleaved	41.3	3.7	6.7	89.3	6.0	7.6	43.1	5.1	8.7	27.9	12.8	15.0	34.3	6.9	9.5
	Labeled	40.5	3.2	6.6	87.0	6.2	7.6	42.6	5.3	8.8	27.1	13.2	15.2	33.4	7.0	9.6
	Multi-task Labeled	41.7	3.5	6.6	89.7	6.1	7.7	43.5	5.0	8.6	27.8	12.6	14.8	34.3	6.8	9.4
	Weighted	46.1	3.6	6.6	102.5	4.6	6.7	49.4	3.8	6.6	32.8	11.4	15.0	39.9	5.8	8.7
GPT-2	PersonaChat (single)	478.8	4.9	6.7	159.6	5.5	6.7	264.7	5.1	7.7	19.6	14.1	16.2	44.7	7.3	9.3
	Concatenated	392.8	5.0	6.9	110.7	5.8	7.0	199.2	5.8	8.5	19.0	13.9	16.0	40.1	7.6	9.6
	Interleaved	26.6	4.3	7.4	54.8	5.8	7.4	28.1	5.7	9.2	19.2	14.0	16.1	23.7	7.4	10.0
	Labeled	26.5	4.2	7.3	54.1	5.9	7.6	27.7	5.7	9.2	18.9	14.1	16.3	23.5	7.5	10.1
	Multi-task Labeled	26.9	4.1	7.2	55.4	5.8	7.5	38.5	5.8	9.4	20.7	14.0	16.1	25.1	7.4	10.1
	Weighted	29.6	4.3	7.5	64.1	5.1	7.4	44.1	4.1	7.0	23.4	13.0	15.7	28.4	6.6	9.4

Table 9: Perplexity, BLEU (%) and F1 (%) scores for baselines and proposed methods fine-tuned on 4 corpora (**with** stop words). BLEU is from NLTK sentence BLEU

Models of labeled, multi-task labeled and weighted learning do not have the best hyper-parameters, but the same hyper-parameters as the base models. Their perplexity is slightly worse than it should be.

The results of the single corpus PersonaChat trained with the LSTM model confirm our concern on a small fine-tuning corpus. The LSTM model is pre-trained on OSDB, Twitter and Ubuntu; however, the performance for the 3 corpora greatly decreases after fine-tuning.

The automatic evaluation with stop words is not good for measuring relevance, since stop words are taken too much into account. See BLEU and F1 scores of PersonChat (single) and weighted learning as an example. Models trained on PersonaChat (single) cannot answer Ubuntu technical questions **at all**, yet they receive better scores than weighted learning. But once the stop words are removed, the scores of weighted learning surplus PersonaChat (single) a lot.

D Additional Results of automatic evaluation without stop words

Model	Corpus / Method	Test set														
		OSDB			Twitter			Ubuntu			PersonaChat			Overall		
		BLEU	ROUGE	DF-F1	BLEU	ROUGE	DF-F1	BLEU	ROUGE	F1	BEU	ROUGE	DF-F1	BLEU	ROUGE	DF-F1
LSTM	PersonaChat (single)	5.2	8.1	6.2	5.7	8.9	5.0	4.5	7.7	4.8	34.2	45.8	44.6	12.4	17.6	15.2
	Concatenated	4.5	7.2	5.6	7.4	11.4	8.8	11.6	20.1	17.4	34.6	46.4	44.2	14.5	21.3	19.0
	Interleaved	6.5	11.7	9.9	8.6	14.9	12.6	17.1	29.6	28.4	32.4	44.3	43.2	16.1	25.1	23.5
	Labeled	6.2	11.3	9.7	9.1	15.5	12.6	18.1	31.6	30.7	33.5	45.4	43.8	16.7	26.0	24.2
	Multi-task Labeled	6.2	11.1	9.5	8.4	14.5	11.7	16.0	28.3	27.2	31.5	42.7	41.9	15.5	24.2	22.6
	Weighted	7.6	13.4	12.2	7.6	15.6	18.7	24.2	41.2	44.1	33.2	47.1	46.9	18.2	29.3	30.5
GPT-2	PersonaChat (single)	7.1	10.8	9.2	8.7	13.9	10.5	8.8	15.5	12.2	45.0	58.8	56.8	17.4	24.7	22.2
	Concatenated	8.4	12.6	11.0	10.8	17.2	13.7	13.4	22.4	23.3	43.0	56.3	55.7	18.9	27.1	25.9
	Interleaved	14.0	22.3	21.3	12.2	20.1	19.3	25.8	44.3	48.3	44.2	57.6	58.0	24.0	36.1	36.7
	Labeled	13.6	21.6	20.5	13.1	21.1	20.3	25.8	45.0	49.6	45.1	59.0	59.6	24.4	36.7	37.5
	Multi-task Labeled	13.4	21.4	20.4	12.7	20.6	20.1	25.4	43.8	47.6	44.0	57.3	57.4	23.9	35.8	36.4
	Weighted	14.5	23.4	23.4	11.9	22.3	25.2	29.2	48.6	52.5	42.4	55.8	57.6	24.5	37.5	39.7

Table 10: BLEU (%), ROUGE (%) and DF-F1 (%) scores for baselines and proposed methods fine-tuned on 4 corpora (**without** stop words). DF-F1 is ROUGE F1 weighted by test-set α DF

E Full results of αDF for generated responses from multiple corpora

Model	Corpus / Method	Test set: OSDB							
		OSDB		Twitter		Ubuntu		PersonaChat	
		Train	Test	Train	Test	Train	Test	Train	Test
Test Set (Standard Score)		7.01	9.66	3.75	3.75	2.82	2.86	3.59	3.75
LSTM	PersonaChat (single)	2.92	3.40	2.40	2.82	2.27	2.51	9.18	9.91
	Concatenated	2.92	3.35	2.49	2.94	2.41	2.71	7.65	8.55
	Interleaved	3.88	4.13	2.45	2.54	2.89	2.87	4.98	5.31
	Labeled	3.94	4.16	2.37	2.44	2.71	2.70	5.01	5.34
	Multi-task Labeled Weighted	3.78	4.02	2.41	2.49	2.91	2.88	5.02	5.36
		5.60	6.29	2.65	2.84	2.89	2.84	4.14	4.47
GPT-2	PersonaChat (single)	2.76	3.15	2.30	2.66	2.24	2.51	10.53	11.09
	Concatenated	3.07	3.59	2.52	2.96	2.30	2.55	8.75	9.35
	Interleaved	4.86	5.78	2.63	2.67	2.69	2.66	4.77	5.04
	Labeled	4.86	5.77	2.61	2.66	2.67	2.64	4.76	5.04
	Multi-task Labeled	4.81	5.70	2.60	2.64	2.69	2.65	4.83	5.1
	Weighted	6.02	7.46	2.71	2.83	2.47	2.48	4.12	4.38

(a) αDF_d scores for generated responses from OSDB

Model	Corpus / Method	Test set: Twitter							
		OSDB		Twitter		Ubuntu		PersonaChat	
		Train	Test	Train	Test	Train	Test	Train	Test
Test Set (Standard Score)		3.97	4.07	9.07	11.01	3.24	3.40	3.64	3.80
LSTM	PersonaChat (single)	2.79	3.21	2.78	3.36	2.35	2.59	8.60	9.18
	Concatenated	2.62	3.12	3.55	4.31	2.30	2.71	7.97	8.69
	Interleaved	3.28	3.68	4.66	4.95	3.11	3.34	4.11	4.51
	Labeled	3.30	3.68	4.97	5.27	3.00	3.24	3.89	4.26
	Multi-task Labeled	3.31	3.68	4.47	4.73	3.14	3.36	4.08	4.49
	Weighted	3.10	3.62	9.92	10.10	2.79	3.01	3.79	4.30
GPT-2	PersonaChat (single)	2.74	3.04	2.87	3.33	2.45	2.66	9.47	9.77
	Concatenated	2.87	3.28	3.32	3.94	2.41	2.65	8.21	8.68
	Interleaved	3.42	3.67	4.59	5.08	3.05	3.13	4.39	4.68
	Labeled	3.48	3.74	4.66	5.16	3.08	3.19	4.06	4.35
	Multi-task Labeled	3.41	3.66	4.63	5.11	3.08	3.15	4.37	4.65
	Weighted	3.58	4.01	8.13	8.84	2.59	2.79	3.68	4.07

(b) αDF_d scores for generated responses from Twitter

Model	Corpus / Method	Test set: Ubuntu							
		OSDB		Twitter		Ubuntu		PersonaChat	
		Train	Test	Train	Test	Train	Test	Train	Test
Test Set (Standard Score)		2.69	2.74	2.96	2.85	19.36	23.20	2.67	2.78
LSTM	PersonaChat (single)	2.71	3.28	2.41	2.89	2.74	3.06	8.55	9.09
	Concatenated	2.61	2.89	2.27	2.53	7.60	7.74	5.59	5.99
	Interleaved	2.91	3.19	2.30	2.36	11.78	11.27	3.70	4.01
	Labeled	3.03	3.38	2.28	2.36	12.46	11.75	3.45	3.75
	Multi-task Labeled	2.91	3.17	2.30	2.35	11.19	10.72	3.77	4.09
	Weighted	2.16	2.84	2.05	2.16	27.73	25.42	2.68	3.01
GPT-2	PersonaChat (single)	2.60	2.85	2.31	2.64	4.12	4.64	8.27	8.42
	Concatenated	2.67	3.03	2.45	2.82	6.54	7.10	7.04	7.37
	Interleaved	2.73	3.05	2.22	2.37	15.67	16.02	3.08	3.41
	Labeled	2.68	3.03	2.17	2.35	16.73	17.02	2.90	3.24
	Multi-task Labeled	2.73	3.06	2.22	2.37	15.45	15.78	3.12	3.44
	Weighted	2.26	2.56	2.16	2.28	25.73	24.42	2.37	2.60

(c) αDF_d scores for generated responses from Ubuntu

Model	Corpus / Method	Test set: PersonaChat							
		OSDB		Twitter		Ubuntu		PersonaChat	
		Train	Test	Train	Test	Train	Test	Train	Test
Test Set (Standard Score)		3.32	3.23	3.18	3.04	2.67	2.69	9.45	12.00
LSTM	PersonaChat (single)	2.59	3.02	2.31	2.73	2.15	2.35	11.86	12.62
	Concatenated	2.47	2.84	2.29	2.76	2.06	2.33	12.52	13.61
	Interleaved	2.57	2.92	2.30	2.71	2.17	2.45	11.48	12.52
	Labeled	2.51	2.88	2.27	2.68	2.08	2.36	12.06	13.11
	Multi-task Labeled Weighted	2.55	2.91	2.29	2.74	2.15	2.43	11.45	12.59
		2.21	2.44	2.13	2.41	2.04	2.18	17.65	18.31
GPT-2	PersonaChat (single)	2.54	2.79	2.28	2.59	2.12	2.34	12.85	13.74
	Concatenated	2.58	2.99	2.41	2.79	2.16	2.39	12.08	12.99
	Interleaved	2.64	2.89	2.37	2.65	2.23	2.42	12.13	12.87
	Labeled	2.57	2.84	2.32	2.62	2.16	2.37	12.37	13.10
	Multi-task Labeled	2.65	2.90	2.37	2.65	2.22	2.42	12.14	12.86
		2.39	2.63	2.27	2.52	2.02	2.17	15.96	17.07

(d) αDF_d scores for generated responses from PersonaChat

Table 11: Full results of αDF_d scores for generated responses from multiple corpora

F Example of human evaluation system

View instructions

Instruction:

There is a context taken from **technical forum chatlogs**.

From the 6 candidate responses, you need to choose the one(s) that is/are the most **RELEVANT/IRRELEVANT** to the context.

Note! You are asked to choose the most **RELEVANT/IRRELEVANT** one(s), not the best/worst one(s).

*You do **NOT** need to consider if the **facts** in the responses are **correct or not**.

Know any mac people in here?

1. No, I am not into them. I like to eat them.
2. I do not. I'm a student.
3. I'm in the same boat.
4. I'm not sure, I'm not sure if I can help you.
5. I'm in the city. I'm in a lot of places.
6. I don't know, I just installed ubuntu on a laptop.

Which response(s) is/are the most RELEVANT to the context?

1. No, I am not into them. I like to eat them.
2. I do not. I'm a student.
3. I'm in the same boat.
4. I'm not sure, I'm not sure if I can help you.
5. I'm in the city. I'm in a lot of places.
6. I don't know, I just installed ubuntu on a laptop.

Which response(s) is/are the most IRRELEVANT to the context?

1. No, I am not into them. I like to eat them.
2. I do not. I'm a student.
3. I'm in the same boat.
4. I'm not sure, I'm not sure if I can help you.
5. I'm in the city. I'm in a lot of places.
6. I don't know, I just installed ubuntu on a laptop.

Submit

Figure 3: Human evaluation system for Ubuntu contexts

G Examples of generated responses

OSDB Context: You know the main reason parents kill kids?	Ubuntu Context: What are you going to do on the remote system exactly?
PersonaChat (single): I am not sure. I am a vegan. Concatenated: I know. I've a daughter that just turned 18. Interleaved: Because they're afraid of what they're gonna do. Labeled: Because they're not going to be able to handle it. Multi-Task Labeled: Because they're afraid of what they're gonna do. Weighted: Because they're afraid of the consequences.	PersonaChat (single): I am going to be a pilot. I am going to fly planes. Concatenated: I am going to go to the beach. Interleaved: I'm going to install the latest version of Ubuntu. Labeled: I'm going to install a new kernel. Multi-Task Labeled: I'm going to install the latest version of Ubuntu. Weighted: I'm trying to connect to the server.
OSDB Context: The big banks will fail, triggering a domino effect.	Ubuntu Context: Know any mac people in here?
PersonaChat (single): I am not sure yet. I am sure they will fail. Concatenated: I'm not sure. I'm not sure what the big banks are doing. Interleaved: What? Labeled: I'm not sure what you're talking about. Multi-Task Labeled: I'm not sure what you mean. Weighted: What's going on?	PersonaChat (single): No, I am not into them. I like to eat them. Concatenated: I do not. I'm a student. Interleaved: I'm in the same boat. Labeled: I'm not sure, I'm not sure if I can help you. Multi-Task Labeled: I'm in the city. I'm in a lot of places. Weighted: I don't know, I just installed ubuntu on a laptop.
OSDB Context: Rigor mortis has yet to set in. His core body temperature remains at 98 degrees. His skin color, lividity is healthy.	Ubuntu Context: Type sudo name of application (in short)
PersonaChat (single): I'm not sure if I like him or not. I like to eat meat. Concatenated: That is good. I like to cook. I am a chef. Interleaved: He is a very good doctor. Labeled: He's a good guy. Multi-Task Labeled: He is a man of great character. Weighted: He's been in the hospital for two weeks.	PersonaChat (single): I like to play games on my computer. Concatenated: I did not see that. I am a programmer. Interleaved: I have a root shell, but I can't login. Labeled: I have a root shell. Multi-Task Labeled: I don't know what that is. Weighted: It is Sudo apt-get install xubuntu-desktop.
Twitter Context: Hillary ran away from rain today. Trump is back on stage minutes after assassination attempt.	PersonaChat Context: That must be hard. I like a beer after driving my nissan pathfinder.
PersonaChat (single): I'm not sure what that means. Concatenated: That is so sad. I hope she gets better soon. Interleaved: I'm so sorry. I'm so sorry. Labeled: I'm so sorry. Multi-Task Labeled: I'm not sure what happened to her. Weighted: She's a liar.	PersonaChat (single): I like to go to the movies. I have been to a few. Concatenated: I am a huge gamer, I play dungeons and dragons on my computer. Interleaved: I like to drive my nissan camry. Labeled: I like to go to the movies with my friends. Multi-Task Labeled: I like to drive my prius. Weighted: I love beer too. I love to drink it while I am driving.
Twitter Context: It was gruesome & extremely rare to have two witnesses. It was linked to epstein. Media didn't report it.	PersonaChat Context: A food bank near my house. Do you ever volunteer?
PersonaChat (single): I'm sorry to hear that. I'm a nurse. Concatenated: I'm not sure what that means. I'm a big fan of the show. Interleaved: I'm not sure what happened to the epstein case. Labeled: I'm not sure what the link was, but it was a link to a video. Multi-Task Labeled: I'm not sure what happened to the epstein case. Weighted: The media is biased and biased.	PersonaChat (single): I do, I love to volunteer at the soup kitchen. Concatenated: Yes I do. I love to help people. Interleaved: I do. I have a lot of friends. Labeled: I do. I am a volunteer at the local soup kitchen. Multi-Task Labeled: I do. I love it. I volunteer at the animal shelter. Weighted: I do. I love volunteering at the food pantry.
Twitter Context: This is 100% true, but can we talk about it tomorrow and make argentina the focus today? Something needs to stick.	PersonaChat Context: Yes! My son is in junior high and I just started letting him watch them too.
PersonaChat (single): Yes, I think so. I'm going to go to the gym tomorrow. Concatenated: I'm going to go see a movie. Interleaved: I'm not sure what that's. I'm not sure what that's. Labeled: I'm not sure what that means. I'm not sure what that means. Multi-Task Labeled: I'm not sure what you mean by that. I'm not sure what you mean by that. Weighted: I'm not sure if I can talk about it tomorrow.	PersonaChat (single): I am a mom of two boys. I have a son. Concatenated: I am a mom, I have three kids. Interleaved: I love kids, I have a son who is in high school. Labeled: I bet he is a good kid. I am a teacher. Multi-Task Labeled: I am sorry to hear that. I am a teacher. Weighted: I bet you are a good mom.

Table 12: Responses generated from GPT-2 fine-tuned on 4 corpora with multiple methods