Analyzing the Forgetting Problem in the Pretrain-Finetune Framework for Neural Language Generation Models

Anonymous authors
Paper under double-blind review

Abstract

In this work, we study how the large-scale pretrain-finetune framework changes the behavior of a neural language generator. We focus on the transformer encoder-decoder model for the open-domain dialogue response generation task. We find that after standard fine-tuning, the model *forgets* important language generation skills acquired during large-scale pre-training. We demonstrate the forgetting phenomenon through a detailed behavior analysis from the perspectives of context sensitivity and knowledge transfer. Adopting the concept of data mixing, we propose an intuitive fine-tuning strategy named "mix-review". We find that mix-review effectively regularize the fine-tuning process, and the forgetting problem is largely alleviated. Finally, we discuss interesting behavior of the resulting dialogue model and its implications.

1 Introduction

Large-scale unsupervised pre-training (Peters et al., 2018; Yang et al., 2019; Liu et al., 2019; Devlin et al., 2018; Song et al., 2019) has recently been shown to greatly boost the performance of natural language processing (NLP) models, and has attracted much research interest. Despite its huge success, there is a fundamental question remaining to be answered:

Is there some crucial weakness in the standard NLP pretrain-finetune framework?

In this work, we take the viewpoint of language generation and show that the answer is, to some extent, **yes**. In particular, we find that the key to answer this question is a concept we denote as *data separation*.

Although various unsupervised pre-training strategies have been proposed for better utilization of large-scale text data, on a high level the pretrain-finetune framework can be viewed as a simple two-stage procedure: (1) use large-scale text data to pre-train the model, and (2) use target task data to fine-tune the model. Data separation refers to (almost) zero-overlapping data usage of the two stages.

In this work we study the pretrain-finetune framework from the viewpoint of neural language generation (NLG). In particular, we focus on the open-domain dialogue response task, for the following reasons: (1) There is high similarity between the target dialogue response task (conditional NLG) and the pre-training language modeling (LM) objective, so we expect that language generation skills learnt during pre-training can be well transferred to the down-stream target task. (2) The sequence-to-sequence (seq2seq) nature of the model allows us to characterize the model's generation behavior in various ways (e.g. context sensitivity).

We briefly summarize our contributions as follows. To study how pretrain-finetuning changes the model's behavior, we conduct a behavior analysis from the perspectives of context sensitivity and knowledge transfer. Our main finding is that in the fine-tuning stage, data separation causes the model to *forget* important language generation skills acquired during pre-training. Motivated by this analysis, we adopt the concept of *data mixing* and propose a mix-review fine-tuning strategy, where we combine the pre-training and fine-tuning objective. We find that mix-review effectively regularize the fine-tuning process, and the forgetting problem is largely alleviated. Finally, we demonstrate and discuss interesting behavior of the resulting dialogue model and its implications.

Dialogue	
Context Input: what did you do yesterday ? `<eou>` i watched the avengers movie .	
Target Output: wow ! i am crazy about iron man !	
Next-sentence Pre-training	
Context Input: the avengers are super hot currently . `<eou>` the next movie will be on in April .	
Target Output: fans are talking about what iron man will do on the internet .	
MASS Pre-training	
Context Input: fans are talking about `<MASK>` `<MASK>` `<MASK>` will do on the internet .	
Target Output: what iron man	

Table 1: Illustrations of input-output pairs for typical dialogue response training, next-sentence pre-training, or MASS pre-training.

2 TRAINING OBJECTIVE FOR SEQ2SEQ TASKS

End-to-end dialogue response generation (Li et al., 2016) can be formulated as a sequence-to-sequence (seq2seq) task: given a dialogue context (previous utterances), the model is asked to generate a high-quality response. In this work we adopt the encoder-decoder model architecture (Sutskever et al., 2014; Cho et al., 2014; Mikolov et al., 2010), which is widely used in NLG applications like dialogue response generation (Li et al., 2016), machine translation (Luong et al., 2015), etc. In particular, we use the transformer model (Vaswani et al., 2017), which has currently become the most popular encoder-decoder model architecture (Young et al., 2017). We use the same configuration as Vaswani et al. (2017), which has 6 encoder/decoder layers, 16 attention heads, with an embedding dimension of 1024 and a feed-forward dimension of 4096.

During baseline training, the Adam optimizer (Kingma & Ba, 2014) is used to minimize the negative log-likelihood (NLL) of the reference target sentence y given the input sentence x in the data distribution (denoted as P_{data}):

$$\mathcal{L}_{\text{MLE}}(P_{data}; \theta) = E_{(\boldsymbol{x},\boldsymbol{y}) \sim P_{data}}(-\log P_\theta(\boldsymbol{y}|\boldsymbol{x})) = E_{(\boldsymbol{x},\boldsymbol{y}) \sim P_{data}}(-\sum_{t=1}^{m} \log P_\theta(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x})), \quad (1)$$

where $\boldsymbol{y}_{<t}$ refers to $\{y_0, y_1, ..., y_{t-1}\}$, in which y_0 is set to a begin-of-sentence token `<BOS>`, and y_m is a end-of-sentence token `<EOS>`. In the dialogue response setting, the input x is a concatenation of previous utterances. We truncate the length of x to be at most 128 words, which typically includes around 6 previous utterances.

Given a trained seq2seq model, to generate a response for some contextual input, one needs to choose a decoding method. Recent research (Holtzman et al., 2019; Radford et al., 2019; Fan et al., 2018) has shown that a strategy called top-k sampling, in which the next word is sampled from the top k most probable choices, is a better choice than the traditional beam-search decoding. Our preliminary experiments (Appendix A) have also verified this claim in the open-domain dialogue response setting. As a result, in this work, unless otherwise mentioned, we use top-k sampling as the default decoding method. In particular, we set k to 30 (we find it to work well in preliminary experiments).

3 THE PRETRAIN-FINETUNE FRAMEWORK

In this section we review the pretrain-finetune framework for encoder-decoder models. More importantly, we discuss the language generation skills the model can acquire during pre-training, and how well they are transferred to the target task. This discussion leads to the proposition of the mix-review fine-tuning strategy.

3.1 PRE-TRAINING

In this work, we consider pre-training the seq2seq model using large-scale unsupervised text data, and afterwards fine-tuning it using target dialogue data. We compare two representative strategies: next-sentence (NS) pre-training and masked sequence-to-sequence (MASS) pre-training (Song et al., 2019). Next-sentence pre-training is a natural extension of GPT-style LM training (Radford et al., 2019; Kiros et al., 2015) for encoder-decoder models. For every sentence in a given training

document, we set the previous sentences as the contextual input, and ask the model to generate the next sentence. We omit the formulation of NS because it is very similar to Equation (1).

Masked sequence-to-sequence pre-training (MASS) can be regarded as an extension of the "BERT" (Devlin et al., 2018) pre-training for encoder-decoder models. For each sentence, a random segment of the sentence is masked, and the model is trained to generate the masked words on the decoder side. We refer readers to Song et al. (2019) for more details.

In Table 1, we illustrate the similarity between NS pre-training and typical dialogue response training. Compared to NS pre-training, MASS has the disadvantage that it focuses on one single sentence at a time. However, the context of multiple previous sentences are very important for dialogue response generation.

There are two important generation capabilities that the model can acquire in the pre-training stage, which will be useful for the target dialogue setting. One is the *acquisition of knowledge* (studied in Section 5.3): the large-scale pre-training text data contains a large amount of knowledge, and can be used to make dialogue responses more informative and engaging (e.g. the model can learn about the "Avengers" movie, and use it as a topic). The other is the *utilization of contextual input* (studied in Section 5.2): as shown by Sankar et al. (2019), the current open-domain dialogue models (without pre-training) are insensitive to contextual input, which gives rise to the generic response problem (Li et al., 2016). In our preliminary experiments with NS pre-training, we find that, similarly to the GPT model (Radford et al., 2019), the pre-trained model has the ability to generate closely related responses given the previous sentences as input. Ideally during fine-tuning, the model can transfer this skill to the target dialogue task.

3.2 THE MIX-REVIEW FINE-TUNING STRATEGY

Although recently a number of pre-training strategies (Peters et al., 2018; Devlin et al., 2018; Song et al., 2019; Yang et al., 2019; Liu et al., 2019) have been proposed for various NLP tasks, the fine-tuning stage remains simple and straightforward: simply fine-tune all parameters with a relatively small learning rate.
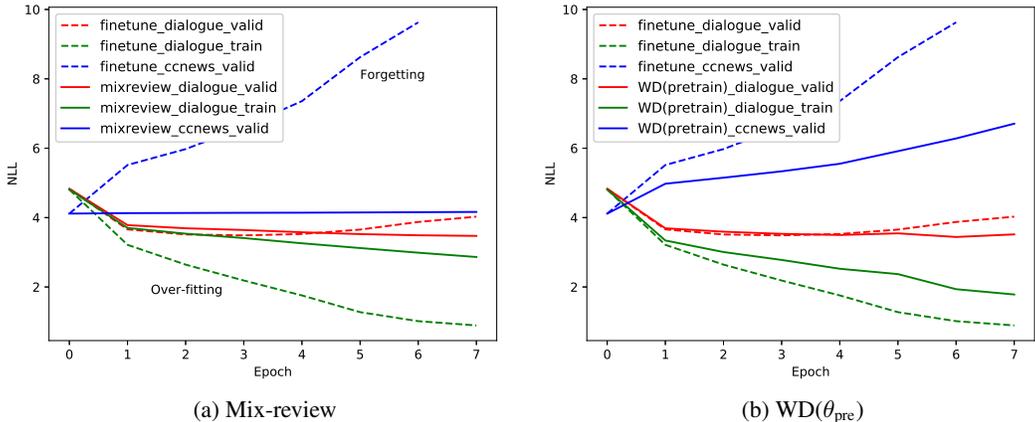

(a) Mix-review   (b) WD(θ_{pre})

Figure 1: Model's performance on different evaluation sets during the fine-tuning stage, for the Dailydialogue data-set (described in Section 4.1).

In Figure 1a, we show the model's negative log-likelihood (NLL) on different evaluation sets during the fine-tuning stage. We identify two potential issues during fine-tuning. (1) **Over-fitting:** The gap between training-set NLL and validation-set NLL increases quickly. (2) **Forgetting:** The performance on the pre-training CCNEWS data (described in Section 4.1) drops drastically. Note that the forgetting phenomenon here is not necessarily "catastrophic" as in the sequential learning case (Atkinson et al., 2018; Robins, 1995), because the goal is to achieve the best performance on the target dialogue data-set, and the model does not need to maintain fidelity to the pre-training data. However, it leads us to suspect that the model has lost some important skills learned during pre-training (verified in Section 5.2 and 5.3).

To address the forgetting phenomenon, we propose a fine-tuning strategy named "mix-review": For each fine-tuning epoch, we mix the target dialogue data with a random subset of the pre-training data. This process introduces two hyper-parameters: *mix-ratio*, which controls how much pre-training data is mixed, and *mix-decay*, which decays the amount of mixed data by each epoch. For example, assume the target dialogue training set has 100k utterances, mix-ratio=4 and mix-decay=0.9, then in the first epoch of mix-review fine-tuning, 400k pre-training utterances will be mixed in, and for the second epoch the amount will be reduced to 360k utterances, etc.

We formulate the mix-review objective as below:

$$\mathcal{L}_{\text{fine-tune}}(P_{\text{target-data}}; \theta) + (\text{mix-ratio} \cdot \text{mix-decay}^{\text{epoch-id}}) \cdot \mathcal{L}_{\text{pre-train}}(P_{\text{pretrain-data}}; \theta) \tag{2}$$

where epoch-id starts from 0. Note that the augmented mixing term can be viewed as a regularization term.

In our experiments, we tune the hyper-parameters (mix-ratio and mix-decay) in the grid of $\{1, 2, 4, 8, 16\} \times \{1, 0.9, 0.8, 0.7, 0.6, 0.5\}$ (using the same learning rate and other hyper-parameters with standard fine-tuning), and report with the best model based on the perplexity (PPL) performance on the validation set of the target task. We find that the performance gain of mix-review is not sensitive to hyper-parameter tuning: A small mix-ratio of 4 typically works well, which means the computational cost of mix-review is comparable to standard fine-tuning.

In Figure 1a, we show the loss curve for mix-review fine-tuning with a mix-ratio of 4 and a mix-decay of 0.7. We observe that the performance on the pre-training CCNEWS data is preserved, which strongly supports the motivation of mix-review. Furthermore, we observe a regularization effect from mix-review (narrowing the gap between training and testing performance).

We compare mix-review with the L_2 regularization (weight decay) toward the pre-trained parameters θ_{pre} (Kirkpatrick et al., 2016; Wiese et al., 2017). We denote it as $\text{WD}(\theta_{\text{pre}})$ and formulate it as follows:

$$\mathcal{L}_{\text{fine-tune}}(P_{\text{target-data}}; \theta) + \lambda \cdot \|\theta - \theta_{\text{pre}}\|_2^2 \tag{3}$$

In our experiments, we tune λ in the set $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ and report with the best model based on PPL on the validation set.

In Figure 1b we show the loss curve for $\text{WD}(\theta_{\text{pre}})$ with $\lambda = 0.1$. We observe that $\text{WD}(\theta_{\text{pre}})$ also has a regularization effect, but it is not as strong as mix-review.

Additionally, we tried the following two basic regularization techniques: (1) Increase the rate of dropout; (2) Freeze the bottom layers of the model during fine-tuning. We find that these two techniques show little or no improvement. We believe the reason is that the transformer is already a well-tuned model (e.g. it features dropout and layer normalization (Lei Ba et al., 2016)).

4 DATA-SETS AND IMPLEMENTATION DETAILS

4.1 DATA-SETS

For pre-training, we use the large-scale CCNEWS data (Bakhtin et al., 2019) which is a de-duplicated subset of the English portion of the CommonCrawl news data-set[1]. The dataset contains news articles published worldwide between September 2016 and February 2019. It has in total around 1 billion sentences or 27 billion words. To be able to complete experiments in a reasonable amount of time, we use the first 10 percent of the CCNEWS data for pre-training, which contains 100 million sentences and 2.7 billion words.

For fine-tuning, three open-domain conversational dialogue data-sets are used: Dailydialogue (1.3 million words) (Li et al., 2017), Switchboard (1.2 million words), and Cornell Movie (Danescu-Niculescu-Mizil & Lee, 2011) (4.5 million words). To save space, we defer the details of the data-sets to Appendix B.

To construct the vocabulary, we learn codes of Byte Pair Encoding (BPE) (Sennrich et al., 2016) from the CCNEWS-100m data with 50k merges. This results in a vocabulary of size 62k. We then apply the same BPE codes to all target dialogue data-sets.

---

[1] http://commoncrawl.org/2016/10/news-dataset-available

Training	Test-PPL/AMT Rating		
	Dailydialogue	**Switchboard**	**Cornell Movie**
Baseline(from scratch)	24.83/(6.323 $\pm$ 0.056)	51.14/(5.269 $\pm$ 0.052)	49.48/(5.844 $\pm$ 0.056)
MASS+finetune	12.78/(6.511 $\pm$ 0.050)	28.41/(5.252 $\pm$ 0.049)	30.25/(5.955 $\pm$ 0.065)
NS+finetune	11.54/(6.515 $\pm$ 0.060)	26.37/(5.332 $\pm$ 0.053)	28.06/(5.932 $\pm$ 0.061)
NS+WD(θ_{pre})	11.19/(6.553 $\pm$ 0.056)	26.25/(5.439 $\pm$ 0.055)	27.80/(5.961 $\pm$ 0.062)
NS+mix-review	**11.07**/(6.577 $\pm$ 0.053)	**25.92**/(5.414 $\pm$ 0.054)	**27.54**/(5.998 $\pm$ 0.055)
Reference	NA/(6.816 $\pm$ 0.052)	NA/(5.630 $\pm$ 0.053)	NA/(6.071 $\pm$ 0.056)

Table 2: Perplexity and AMT-Rating evaluation for different training process on the three dialogue data-sets. The rating scores are the average score of fluency, consistency, and engagingness.

4.2 IMPLEMENTATION

Our code is based on the Fairseq toolkit (Ott et al., 2019). The Adam optimizer (Kingma & Ba, 2014) is used for all experiments. For pre-training of both MASS and NS, we use a mini-batch size of 2048, with the learning rate (LR) set to 0.0001. Following Vaswani et al. (2017), the "inverse square root" LR scheduler with a warm-up stage is used. Pre-training is conducted on 32 GPUs and half-precision (float16) speed-up is used. For both MASS and NS, we stop the pre-training after the CCNEWS data is swept 20 times. Although the perplexity is still improving, we stop the pre-training for practical reasons to control the duration of the experiments. For all our experiments, a dropout rate of 0.1 is applied to the transformer model. We follow Song et al. (2019) for the recommended hyper-parameter setting of MASS (e.g. how to select the mask span).

Fine-tuning (with or without mix-review) is done on 2 GPUs without float16 speed-up. The learning rate is halved when the PPL on the validation set does not improve. In almost all fine-tuning experiments over-fitting is observed, and we do an early-stop when performance on the validation set starts to deteriorate. We tune the learning rate from $\{10^{-3}, 10^{-4}, 10^{-5}\}$, and report the best model based on validation set perplexity.

5 EXPERIMENT RESULTS

In this section, we first present results for the standard dialogue model evaluation. We then conduct a detailed behavior analysis, characterising how different training strategies change the model's behavior. In particular, we aim to answer the crucial question about whether the model forgets precious language generation skills during standard fine-tuning, and more importantly, whether mix-review helps the model remember the skills.

5.1 STANDARD DIALOGUE MODEL EVALUATION

In addition to perplexity, we use the Amazon Mechanical Turk (AMT) platform for human evaluation of different training processes on the three dialogue data-sets. For the AMT rating, each turker is given a dialogue context, and a randomly permuted set of model sample responses. The turker is then asked to rate each sample response according to its fluency, consistency, and engagingness respectively, using an integer score from 1 to 9. The reference response is also rated for comparison. For each data-model pair, we collect 2,500 ratings. The results are shown in Table 2. To remove bias among annotators, we use the bayesian inference code from Kulikov et al. (2018), and report calibrated mean and standard deviation. Since we use top-k sampling, the BLEU score is not directly suitable for our setting (Liu et al., 2016).

We first observe the huge improvement in perplexity (larger than 40%) for the pre-trained models comparing to the baseline models trained from scratch. Comparing to MASS, the NS pre-training has more than 7% relative improvement. As discussed in Section 3.1, this confirms our earlier discussion that the model pre-trained by NS better utilizes contextual input (which is further verified in Section 5.2). Based on this observation, we focus our analysis below on the NS pre-training.

Comparing to standard fine-tuning, mix-review further gives solid improvement. The gain is due to its strong regularization effect (which we study in the next two sections). However, the performance gap between mix-review and WD(θ_{pre}) is not significant. We conjecture that mix-review could be too

---

Input: i 'd like a beer . what kind of beer do you have ?
   Baseline: i 'm not sure . could you recommend some ? i 'm not sure .
   MASS: we have various kinds of beer , beer and soft drinks .
   NS: we have a beer , but we don ' t have a lager and a black beer .
   WD($\theta_{\mathbf{pre}}$): i ' d like a beer of medium pace .
   Mix-review: we have a malty and belgian ale . it is a little special .
Input: draft beer and cans beer . which do you prefer ?
   Baseline: well , two large fruit cakes .
   MASS: how about a soft beer ? it ' s very good .
   NS: a bottle of beer with some meat , sir .
   WD($\theta_{\mathbf{pre}}$): do you have any beer with a clear finish ?
   Mix-review: can you make a can , please ?

Table 3: Samples of different models on the Dailydialogue test-set. The samples are consecutive (input of the next sample is the reference response for the previous one). More samples are given in Appendix C.

Model(Data-set)	PPL(**normal**)	PPL(**word-shuffle**)	PPL(**word-drop**)
NS Pre-trained(CCNEWS)	17.33	36.56(**+110.96%**)	35.56(**+105.19%**)
Baseline(Dailydialogue)	24.83	27.87(+12.2%)	31.87(+28.3%)
MASS+finetune(Dailydialogue)	12.78	15.85(+24.0%)	18.13(+41.8%)
NS+finetune(Dailydialogue)	11.54	16.30(+41.2%)	19.01(+64.7%)
NS+WD(θ_{pre})(Dailydialogue)	11.19	14.16(+26.5%)	16.37(+46.2%)
NS+Mix-review(Dailydialogue)	11.07	17.81(**+60.8%**)	23.05(**+108.2%**)

Table 4: The model's PPL performance when word-shuffle or word-drop is applied to the context input. On the left we describe what training process is used and on which test set is PPL evaluated. Note that MASS/NS refers to MASS/NS pre-training with standard fine-tuning. To save space, the results on Switchboard and Cornell Movie data-sets are deferred to Appendix D.

"aggressive" in regularizing the model's generative behavior, and more sophisticated regularization techniques are worth investigating. We revisit this discussion in Section 6.

In Table 3 we compare samples from different models on the Dailydialogue test-set. Compared to the baseline model, we find that the pre-trained model's responses are more related to the context. For example, in the second response, the baseline model talks about "fruit cake", while the pre-trained models talk about beer, which is the main subject of this conversation. The samples from mix-review are interesting in that the model is able to describe beer with sophisticated words like "belgian ale" or "medium-batch", which we attribute to knowledge obtained during pre-training (the phrases "belgian ale" or "medium-batch" do not exist in the Dailydialogue training data).

5.2 BEHAVIOR ANALYSIS: CONTEXT SENSITIVITY

The sensitivity to context is an important property for NLG models. However, as shown by Sankar et al. (2019), dialogue response models trained from scratch typically are not sensitive to artificial distortion in the context input, showing the models have poor utilization of dialogue context. In this section, we repeat their experiments with pre-trained dialogue models.

Following Sankar et al. (2019), we use two methods to distort the context input:

- **word-drop:** We randomly drop 30% of the words in the context input.
- **word-shuffle:** We randomly shuffle the words in the context input.

We use the relative drop in test-set perplexity to quantify the sensitivity. The results are presented in Table 4, where the result of the pre-trained model is also included. First, we observe the baseline model trained from scratch is relatively insensitive to context, which agrees well with Sankar et al. (2019). The model with the standard pretrain-finetune process is much more sensitive, showing that pre-training effectively changes the model's behavior. Comparing to MASS, the NS pre-trained model has better utlization of context, which explains its superior performance (in Section 5.1).

News-style Triggers	Dialogue-style Triggers
now, some opinions about pokemon .	what you do think about pokemon ?
let me tell you about pokemon .	please tell me about pokemon .
here's some news about pokemon .	do you have news about pokemon ?

Reference Description: Pokemon first took the world by storm in the mid-90s, doing so once again this year with the release of Pokemon Go.
NS Pre-trained: the game , titled pokemon go : pocket camp , can be played in person ...
Standard Fine-tuned: it 's a new game that can be played with kids .
WD(θ_{pre}): pokemon go , it 's a type of game that only exists in the us .
Mix-review: pokemon go is a popular mobile game , where you 're expected to catch pokemon .

Reference Description: Deadpool: The wisecracking antihero, played by Ryan Reynolds in a movie of the same name, became the highest-grossing R-rated film of all time.
NS Pre-trained: ryan reynolds teased his upcoming movie as the character of deadpool .
Standard Fine-tuned: it 's a popular movie .
WD(θ_{pre}): yes , i really like him . he is a very funny character .
Mix-review: ryan reynolds .

Table 5: Example of trigger inputs for the knowledge term "pokemon". Followed by reference description and model samples for "pokemon" and "deadpool". Note that the pre-trained model's sample is from news-style triggers, and the other samples are from dialogue-style triggers.

	Dailydialogue		Switchboard	
Model	Dialogue Triggers	News Triggers	Dialogue Triggers	News Triggers
NS Pre-trained	0.245/0.089	**0.347/0.153**	0.245/0.089	**0.347/0.153**
Baseline	0.124/0.007	0.101/0.004	0.032/0.0003	0.046/0.002
NS+finetune	0.162/0.047	0.158/0.046	0.187/0.052	0.170/0.044
NS+WD(θ_{pre})	0.226/0.080	0.235/0.085	0.203/0.070	0.204/0.060
NS+Mix-review	**0.261/0.108**	0.322 /0.135	**0.223/0.079**	0.341/0.151

Table 6: Average BLEU-2/BLEU-3 scores for the model's samples w.r.t. the reference description. We highlight the pre-trained model's performance for news triggers and the performance of the best model fine-tuned with dialogue data for dialogue triggers. The results on Cornell Movie data-set is deferred to Appendix D.

Somewhat surprisingly, the NS pre-trained dialogue models are much less sensitive to context input than the pre-trained model without fine-tuning. This verifies our worry in Section 3.2 that the model is forgetting some important generation skill during standard fine-tuning. Further, we find that the mix-review fine-tuning strategy can effectively alleviate this problem: Its sensitivity is much greater than that of standard fine-tuning, and is close to the pre-trained model.

5.3 BEHAVIOR ANALYSIS: KNOWLEDGE TRANSFER

As argued in Section 3.1, ideally the model can acquire "knowledge" from the large-scale pre-training data, which will be useful for the downstream open-domain dialogue task. In this section, we design a process to quantify how much knowledge the model has, and use it to monitor how the pretrain-finetune framework changes the model's behavior.

Since the pre-training CCNEWS data is in the public news domain, we expect the model to have knowledge about "big news". So, we utilize the Google trend data of the year 2016,[2] which contains 365 trending terms (e.g. iPhone 7, Deadpool, etc.), and its corresponding description.

To query whether the model has knowledge of a certain term, we design three news-style and three dialogue-style "trigger templates" to trigger the model to generate responses related to the knowledge term. We collect 10 samples for each trigger (30 samples from news/dialogue-style triggers for each term), then we compute BLEU score of generated samples against the reference descriptions. We show some examples of trigger inputs in Table 5.

---

[2]https://www.google.com/intl/en-US/trends/2016records/

Dailydialogue	Knowledge	Fluency	Consistency	Engagingness
NS+finetune	2.829 ± 0.061	5.359 ± 0.067	4.288 ± 0.067	3.848 ± 0.051
NS+WD(θ_{pre})	3.189 ± 0.067	5.671 ± 0.059	4.605 ± 0.064	4.184 ± 0.059
NS+Mix-review	$\mathbf{3.401 \pm 0.051}$	5.692 ± 0.060	$\mathbf{4.755 \pm 0.068}$	$\mathbf{4.272 \pm 0.056}$

Table 7: AMT rating scores (calibrated mean and standard deviation) for multi-turn dialogue evaluation.

The BLEU scores are shown in Table 6. Note that we should compare the pre-trained model's scores for the news triggers with the other dialogue models' scores for dialogue triggers. We first observe for the pre-trained model, the news-style triggers can get much more relevant output than the dialogue-style triggers. This matches our intuition because the pre-trained model is trained with news data. Although the fine-tuned model is more knowledgeable than the baseline model, its score is much lower than the pre-trained model. Similar to the case of context sensitivity (Section 5.2), this again demonstrates the forgetting problem of the standard fine-tuning.

We find that mix-review and WD(θ_{pre}) can effectively retain the knowledge acquired during pre-training, giving a much higher BLEU score than the standard fine-tuned model. Mix-review shows higher BLEU scores than WD(θ_{pre}), demonstrating its superiority in facilitating knowledge retention. We showcase samples from different models in Table 5. To save space, we manually select and show the most related sample out of the 30 samples for each knowledge term. The observations agree with the quantitative results: the standard fine-tuning loses the detailed information about the knowledge term, and mix-review helps the model retain it. More importantly, the model is able to express the knowledge in a dialogue context.

To further investigate the ability of our model to conduct knowledgeable dialogues with users, we use the ParlAI[3] platform to conduct multi-turn dialogue evaluation. For each session, the user will be assigned a random knowledge term from Google Trend, and have a 8-turn dialogue with the model under that topic. Ratings from around 600 dialogues are collected for each model, and are reported in Table 7. In this evaluation we use the models fine-tuned on the Dailydialogue data, because the nature of that data-set is closet to online chit-chat. It is observed that the model trained mix-review significantly outperforms WD(θ_{pre}) on knowledge, consistency and engagingness, which agrees well with the results in Table 6 and Table 4. We include some dialogue examples in Appendix C, and Table 8 also has some related examples.
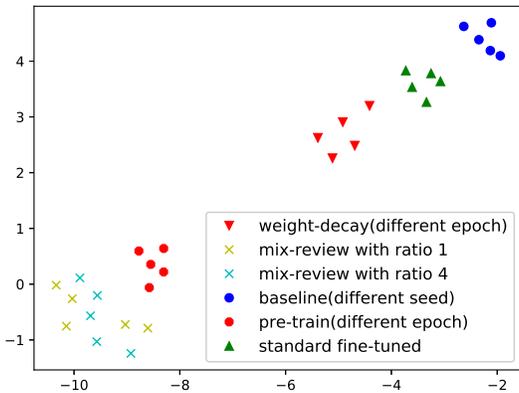
6 IMPLICATIONS AND DISCUSSION

Figure 2: UMAP projection of checkpoints from different training processes.

---

[3]https://parl.ai/

User: have you been to china ?	**User:** who is mark zukerberg ?
Model: yes , i have .	**Model:** he is the ceo of facebook .
User: which city did you go ?	**User:** what is that company about ?
Model: shanghai	**Model:** it is a social media company .
User: how do you like it there ?	**User:** do you like him ?
Model: i like it very much	**Model:** i think he is a good leader .
User: i think the weather is not so good .	**User:** my friend works at facebook .
Model: but it is nice in spring and fall .	**Model:** what does your friend do there ?

User: when is the independence day ? **Best-of-10:** it 's july fourth.
User: is coffee good for health ? **Best-of-10:** it may help reduce your risk of certain diseases.
User: should i do a phd or go to work ? **Best-of-10:** a phd is not enough if you want to learn something new .

User: oh shut up you . **Worst-of-10:** shut up !
User: you 're the problem . **Worst-of-10:** i hate you .
User: i will kill . **Worst-of-10:** i will kill you .

Table 8: Multi-turn and single-turn examples of the model trained by mix-review on Dailydialogue data. The single-turn examples involve light cherry-picking.

In Figure 2, we show the UMAP (McInnes et al., 2018) projection for the models in the function space (Erhan et al., 2010). The details and extra plots are given in Appendix E. We observe that the standard fine-tuned models are not close to the cluster of the pre-trained models, which suggests the models' generative behavior is substantially different from the pre-trained ones. Mix-review regularizes the fine-tuning process to keep the model's generation behavior close to the pre-trained model. These observations agree with our results in Section 5.2 and 5.3. Figure 2 also suggests potential limitations of mix-review and WD(θ_{pre}): Mix-review could be too "aggressive" and does not put enough attention on the target task. On the other hand, WD(θ_{pre}) is not strong enough in regularizing the model's generative behavior (discussed further in Appendix E).

From the viewpoint of the open-domain dialogue task, the sensitivity to dialogue context and the ability to transfer knowledge from pre-training opens the possibility of a data-driven **knowledgable** chat-bot. In Table 8, we show multi-turn and single-turn interaction examples with the model trained by mix-review. For demonstration purpose, we manually select the most interesting response out of 10 samples from the model for the single-turn examples. We observe that the model is able to return interesting responses with the knowledge it acquires from pre-training. More interestingly, it has developed its own "opinions" and is able to give advice to the user.

Finally, we discuss the malicious response problem for open-domain dialogue models. As shown by He & Glass (2019a), it is relatively difficult to trigger the dialogue models trained from scratch to output malicious responses (note that the conversations from the Dailydialogue data tend to be very polite). However, as shown in Table 8, the pre-trained models are easily triggered to respond in a malicious way when "provoked". This is because compared to the baseline models, the pre-trained models are more sensitive to the contextual input, making them easier to manipulate. This makes the malicious response problem a more relevant issue to solve (He & Glass, 2019b).

7 RELATED WORKS

Forgetting As discussed in Section 3.2, in contrast to the "catastrophic forgetting" problem in sequential learning (Atkinson et al., 2018; Robins, 1995; Riemer et al., 2017), the performance drop on pre-training data is not necessarily bad for the NLP pretrain-finetune framework. In Section 5.2 and 5.3, we confirm the "forgetting" of important language generation skills during standard fine-tuning. The proposed mix-review strategy is similar to the *pseudo-rehearsal* algorithm in sequential learning (Robins, 1995), with the difference being that we assume we still have access to the pre-training data. Mix-review can also be viewed as a form of multi-task learning (Li et al., 2019), which has been shown to be useful in neural machine translation (NMT) (Niehues & Cho, 2017), speech recognition (Toshniwal et al., 2017), optical character recognition (OCR) (Liao et al., 2019), etc. However, these works mostly focus on supervised tasks. To the best of our knowledge, this is the first work to analyze the forgetting problem for NLG models under the unsupervised pretrain-finetune work, and address it using the concept of data mixing.

Pre-training for NLG Models Unsupervised pre-training for NLG models has recently received much research attention (Wolf et al., 2019; Mehri et al., 2019; Song et al., 2019; Devlin et al., 2018), but how pre-training changes the behavior of a neural language generator is poorly understood. Several studies have shown that large-scale training teaches LM common-sense knowledge (Petroni et al., 2019; Trinh & Le, 2019), in which the captured knowledge is quantified by a cloze-style test. On the other hand, knowledge-grounded chat-bots (Liu et al., 2018; Zhu et al., 2017) have been an important topic for dialogue models. These studies usually involve additional retrieval modules to provide the model with relevant information. Unlike these works, we study whether fine-tuning *preserves* knowledge gained during large-scale pre-training.

8 CONCLUSION

In this work, we analyze forgetting problem for the standard NLP pretrain-finetune framework in the viewpoint of language generation. We adopt the concept of "data mixing" and propose the mix-review fine-tuning strategy. We demonstrate that mix-review can effectively help the model *remember* important generation skills learned during pre-training.

Through a detailed behavior analysis, we find that under the surface of the performance boost for standard metrics, large-scale pre-training changes the model's generative behavior in various profound ways (e.g. context sensitivity). More importantly, the behavior change is influenced by the nature of data itself. For example, we demonstrate that we can discuss news with the resulting dialogue model, even when the fine-tuning data is not about news (Dailydialogue). This opens the exciting possibility of a completely data-driven way to customize a language generator.

REFERENCES

Craig Atkinson, Brendan McCane, Lech Szymanski, and Anthony V. Robins. Pseudo-recursal: Solving the catastrophic forgetting problem in deep neural networks. *CoRR*, abs/1802.03875, 2018. URL http://arxiv.org/abs/1802.03875.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. Real or fake? learning to discriminate machine from human generated text. *CoRR*, abs/1906.03351, 2019. URL http://arxiv.org/abs/1906.03351.

Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.*, 11:625–660, March 2010. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1756006.1756025.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://www.aclweb.org/anthology/P18-1082.

Tianxing He and James Glass. Detecting egregious responses in neural sequence-to-sequence models. In *International Conference on Learning Representations*, 2019a.

Tianxing He and James R. Glass. Negative training for neural dialogue response generation. *CoRR*, abs/1903.02134, 2019b. URL http://arxiv.org/abs/1903.02134.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *CoRR*, abs/1904.09751, 2019. URL http://arxiv.org/abs/1904.09751.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL http://arxiv.org/abs/1412.6980. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114 13:3521–3526, 2016.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 3294–3302. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5950-skip-thought-vectors.pdf.

Ilya Kulikov, Alexander H Miller, Kyunghyun Cho, and Jason Weston. Importance of a search strategy in neural dialogue modelling. *arXiv preprint arXiv:1811.00907*, 2018.

J. Lei Ba, J. R. Kiros, and G. E. Hinton. Layer Normalization. *arXiv e-prints*, July 2016.

Jianquan Li, Xiaokang Liu, Wenpeng Yin, Min Yang, and Liqun Ma. An empirical evaluation of multi-task learning in deep neural networks for natural language processing. *CoRR*, abs/1908.07820, 2019. URL https://arxiv.org/abs/1908.07820.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 110–119, 2016.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1099.

Minghui Liao, Pengyuan Lyu, Minghang He, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *CoRR*, abs/1908.08207, 2019. URL http://arxiv.org/abs/1908.08207.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 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*, pp. 2122–2132. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1230.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1489–1498, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1138. URL https://www.aclweb.org/anthology/P18-1138.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421. Association for Computational Linguistics, 2015. doi: 10.18653/v1/D15-1166.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskénazi. Pretraining methods for dialog context representation learning. *CoRR*, abs/1906.00414, 2019. URL http://arxiv.org/abs/1906.00414.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pp. 1045–1048, 2010.

Jan Niehues and Eunah Cho. Exploiting linguistic resources for neural machine translation using multi-task learning. *CoRR*, abs/1708.00993, 2017. URL http://arxiv.org/abs/1708.00993.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

Fabio Petroni, Tim Rocktschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases?, 2019.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Matthew Riemer, Elham Khabiri, and Richard Goodwin. Representation stability as a regularizer for improved text analytics transfer learning. *CoRR*, abs/1704.03617, 2017. URL http://arxiv.org/abs/1704.03617.

Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7: 123–146, 1995.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher J. Pal, Sarath Chandar, and Yoshua Bengio. Do neural dialog systems use the conversation history effectively? an empirical study. *CoRR*, abs/1906.01603, 2019. URL http://arxiv.org/abs/1906.01603.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*, 2019.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 3104–3112, 2014.

Shubham Toshniwal, Hao Tang, Liang Lu, and Karen Livescu. Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition. *CoRR*, abs/1704.01631, 2017. URL http://arxiv.org/abs/1704.01631.

Trieu H. Trinh and Quoc V. Le. Do language models have common sense?, 2019. URL https://openreview.net/forum?id=rkgfWh0qKX.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. Neural domain adaptation for biomedical question answering. *arXiv preprint arXiv:1706.03610*, 2017.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149, 2019. URL http://arxiv.org/abs/1901.08149.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019. URL http://arxiv.org/abs/1906.08237.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *CoRR*, abs/1708.02709, 2017. URL http://arxiv.org/abs/1708.02709.

Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. Flexible end-to-end dialogue system for knowledge grounded conversation. *CoRR*, abs/1709.04264, 2017. URL http://arxiv.org/abs/1709.04264.

A   BEAM-SEARCH VS. TOP-K SAMPLING

Data-set	Beam Search		Top-30 Sampling	
	Entropy	**Max-ratio**	**Entropy**	**Max-ratio**
Dailydialogue	7.44 8.49	1.7% 1.3%	9.04 10.81	0.6% 0.4%
Switchboard	4.96 5.54	34.9% 27.8%	8.47 10.45	8.4% 7.9%
Cornell	6.10 6.56	10.2% 9.9%	8.76 10.54	1.4% 1.1%

Table 9: Average of diversity metrics for models on the three dialogue data-sets.

To compare beam search with top-k sampling (we set k to 30), we compute diversity metrics for samples from models trained by different procedures (from scratch or pre-trained). In particular, we compute bi-gram and tri-gram entropy, and the ratio of the most frequent response and second most frequent response (denoted as max-ratio) (He & Glass, 2019b). The results are shown in Table 9.

We observe that the responses given by top-k sampling are much more diverse than beam search. Beam search suffers much from the "generic response" problem (Li et al., 2016), for example, 34% of the responses are "um - hum" for Switchboard. Further, in our multi-turn dialogue experiments, beam-search is likely to give repetitive responses. Finally, by manual inspection, we find the sample quality of top-k sampling is not compromised. Due to these observations, we adopt top-k sampling as the main decoding method for this work.

B   DETAILS ON DATA-SETS

Dailydialogue (Li et al., 2017) is a high-quality multi-turn dialog data-set. The language is human-written and less noisy. The dialogues in the data-set reflect our everyday communication and cover various topics about our daily life. The training split has around 11k dialogues (1.3 million words), and both the validation and test splits have 1k dialogues (0.1 million words).

0 Tell me about the tv show Robot Wars	0 tell me about curry
1 it 's about a group of robots fighting in a war.	1 well, he 's one of the best players in the world.
0 Cool! What season is it on?	0 That's cool, what team does he play for?
1 it 's on next month.	1 the golden state warriors.
0 How many years has it been on for?	0 Do you like the Warriors? I am a Toronto Raptors fan
1 it started in 2003.	1 I like the warriors too.
0 What channel does it come on?	0 Tell me what position he plays and is he good?
1 the science channel.	1 he is good at shooting the ball.
0 What season of the show is it?	0 Is he a good passer? Basketball is a team sport
1 it 's on on fridays and saturdays.	1 yes, he is a team player.
0 Who is on the show?	0 Should the warriors trade him and rebuild the team?
1 it 's an american science channel.	1 yes, I think so.
0 No I mean like what actors star on the show?	0 Who should they get? He could bring good draft picks
1 it 's a robot war.	1 I think he 's a good fit for the warriors.

Table 10: Multi-turn dialogue samples with turkers from the model trained by mix-review. 0 represents the turker, and 1 is the model.

The Switchboard Dialogue Act Corpus[4] is a version of the Switchboard Telephone Speech Corpus, which is a collection of two-sided telephone conversations, annotated with utterance-level dialogue acts. In this work we only use the conversation text part of the data, and select 1.1k dialogues for training (181k sentences / 1.2 million words), 50 dialogues for validation and 50 dialogues for testing.

The Cornell Movie Dialogue Corpus[5] (Danescu-Niculescu-Mizil & Lee, 2011) is a collection of movie scripts. In the processing of the data, we simply regard the whole scripts from a movie as a long dialogue. The training split contains 9k dialogues (4.5 million words), and both the validation and test splits have 180 dialogues (85k words).

C MODEL SAMPLES

In Table 10, we show interaction samples where is turker and the model is talking about an assigned topic from Google Trend.

In Table 11, we show more samples from different training procedure, for the three dialogue data-sets.

D SUPPLEMENTARY EXPERIMENT RESULTS

In this section we supplement results that are deferred in the main body due to space limit.

In Table 12 we show Fluency/Consistency/Engagingness scores of the AMT Rating.

In Table 13 we show context sensitivity results for Switchboard and Cornell Movie data-sets.

In Table 14 we show the knowledge transfer results for the Cornell Movie data-set.

E DETAILS AND AUXILIARY PLOTS OF UMAP PROJECTION

For function space projection, the input to UMAP should be the model's output distributions. We collect the model's output distribution on 10k words for the CCNEWS validation set and the Dailydialogue validation set (so it's a concatenation of two long vectors). We use the default hyper-parameter setting of the python implementation of UMAP. The result is shown in Figure 2 in the main body. Note that during pre-training of the CCNEWS data, 20 epochs are one entire data pass. We fine-tune from epoch 100, 200, 300, 400, 500 of the pre-training checkpoints.

---

[4] http://compprag.christopherpotts.net/swda.html
[5] https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html

Dailydialogue

Input: why do you think you are qualified for this position ?

    Baseline: i can tell you more about your company because your company has a working experience with a company and company company .

    NS: i have been working with microsoft for four years . i can make good use of microsoft ' s word pad and excel excel spreadsheet .

    MASS: i think i have a strong sense of responsibility and i have had a high iq score in college . i am applying for some courses of study for some courses if i have a preference .

    Mix-review: my background is in computer engineering and i have a strong commercial acumen in business - related application . i am in charge of all the things i like to do here .

    WD(θ_{pre}): my background , education , experience as an english professor , and the work ethic are strong .

Input: according to your advertisement , you want an experienced software engineer . i think my background meets the requirement of this position .

    Baseline: yes . and you also need to know how to enjoy the work . and , here is a position to develop the company position under the company .

    NS: do you know your qualifications ?

    MASS: that ' s a good answer . i can tell you if we have made a decision this afternoon .

    Mix-review: i see . do you have some experience as an executive engineer ?

    WD(θ_{pre}): in your first year of operating your company , you had an impressive track record of product design , and achieved a high level of technical proficiency .

Switchboard

Input: that 's going to be a really viable program if it continues to be funded at current levels

    Baseline: oh really

    NS: um it 's going to be a lot of fun it certainly is

    MASS: sure sure sure

    Mix-review: well that that 's what i 'm saying too that 's that 's not gonna be an issue

    WD(θ_{pre}): yeah oh sure sure

Input: uh - huh

    Baseline: oh okay

    NS: and uh i- i think there are some areas that if the government decides to close down that could be a real problem

    MASS: and i really feel strongly that we don 't need the space station because we think there 's really no space

    Mix-review: but uh that

    WD(θ_{pre}): i mean it 's pretty i think it 's it 's pretty interesting to me that it 's uh

Cornell Movie

Input: i expect you shortly .

    Baseline: how so ?

    NS: i 'll make a reservation .

    MASS: i 'll come with you , doctor .

    Mix-review: no , i 'm fine

    WD(θ_{pre}): okay .

Input: what shall we do ?

    Baseline: we all don 't know where the hell he is now , but if he talks to me first , i 'll be back .

    NS: nothing .

    MASS: we send a cab ! i said i expect you to be fast .

    Mix-review: david 's not very lucid , but he 's always lucid .

    WD(θ_{pre}): stop him .

Input: tea would be nice .

    Baseline: i have no desire to go to bed .

    NS: but you can 't come here and have tea too ?

    MASS: i 've been expecting you , doctor .

    Mix-review: do you want me to drink some ?

    WD(θ_{pre}): what is it ?

Table 11: Samples of different models on the Dailydialogue/Switchboard/Cornell-Movie test-set. The samples are consecutive (input of the next sample is the reference response for the previous one).

AMT Ratings	Fluency/Consistency/Engagingness		
Model	**Dailydialogue**	**Switchboard**	**Cornell Movie**
Baseline	6.09/5.90/6.23	5.58/5.59/5.60	5.96/5.72/5.96
MASS+finetune	6.32/6.12/6.35	5.62/5.61/5.49	6.02/5.84/6.05
NS+finetune	6.21/6.21/6.39	5.60/5.67/5.66	5.93/5.84/6.05
NS+WD(θ_{pre})	6.32/6.17/6.45	5.70/5.79/5.79	6.00/5.92/6.06
NS+Mix-review	6.35/6.23/6.41	5.71/5.76/5.74	6.03/5.91/6.11
Reference	6.54/6.46/6.70	5.87/6.00/5.98	6.01/6.02/6.28

Table 12: The detailed rating scores from AMT.

Model(Data-set)	PPL(**normal**)	PPL(**word-shuffle**)	PPL(**word-drop**)
NS Pre-trained(CCNEWS)	17.33	36.56(**+110.96%**)	35.56(**+105.19%**)
Baseline(Switchboard)	51.14	53.42(+4.4%)	53.94(+5.4%)
MASS+finetune(Switchboard)	28.41	32.68(+15.0%)	33.91(+19.3%)
NS+finetune(Switchboard)	26.37	30.87(+17.0%)	32.08(+21.6%)
NS+WD(θ_{pre})(Switchboard)	26.25	31.31(+19.2%)	32.89(+25.2%)
NS+Mix-review(Switchboard)	25.92	31.10(**+19.9%**)	33.70(**+30.0%**)
Baseline(Cornell)	49.48	50.22(+1.4%)	50.85(+2.7%)
MASS+finetune(Cornell)	30.25	36.50(+20.6%)	36.36(+20.1%)
NS+finetune(Cornell)	28.06	36.88(+31.4%)	34.47(+22.8%)
NS+WD(θ_{pre})(Cornell)	27.80	37.46(**+34.7%**)	35.10(+26.2%)
NS+Mix-review(Cornell)	27.54	36.94(+34.1%)	37.72(**+36.9%**)

Table 13: The model's PPL performance when word-shuffle or word-drop is applied to the context input. On the left we describe what training process is used and on which test set is PPL evaluated. Note that MASS/NS refers to MASS/NS pre-training with standard fine-tuning.

	Cornell	
Model	Dialogue Triggers	News Triggers
NS Pre-trained	0.245/0.089	**0.347/0.153**
Baseline	0.081/0.003	0.088/0.003
NS+finetune	0.207/0.071	0.207/0.063
NS+WD(θ_{pre})	0.285/0.114	0.202/0.072
NS+Mix-review	**0.396/0.190**	0.212/0.065

Table 14: Average BLEU-2/BLEU-3 scores for the model's samples w.r.t. the reference description. We highlight the pre-trained model's performance for news triggers and the performance of the best model fine-tuned with dialogue data for dialogue triggers.
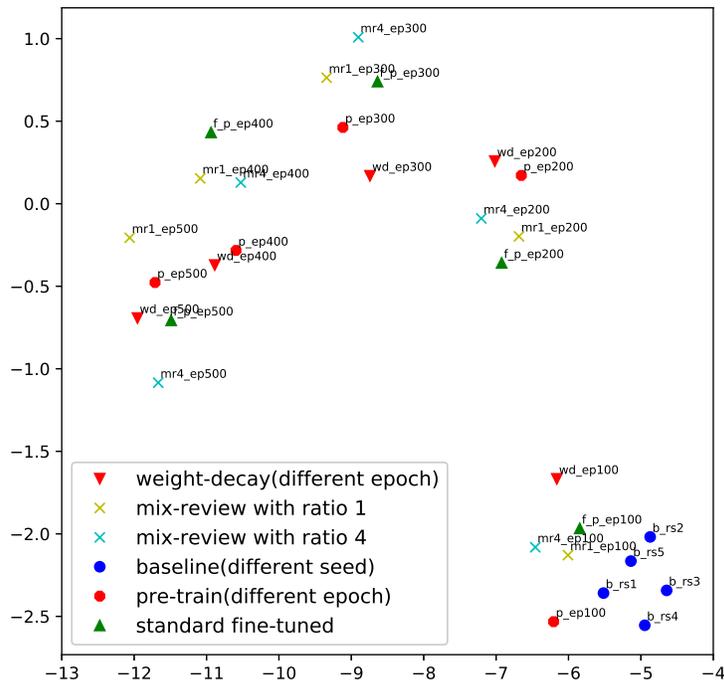
Figure 3: Parameter-space UMAP projection of checkpoints from different training processes.

In Figure 3 we show the parameter space UMAP projection for the same set of models. In this case, the input to UMAP is the concatenation of flattened weight matrices of the transformer model. A key observation is that the fine-tuned models are typically very close to the starting point (pre-trained models). However, as shown in Figure 2, their behavior is very different. This suggests that a parameter-space regularization such as WD(θ_{pre}) could be not very effective for regularizing the model's behavior.