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 $P_{\text{data}}$):

$$L_{\text{MLE}}(P_{\text{data}}; \theta) = E_{(x,y) \sim P_{\text{data}}}(- \log P_{\theta}(y|x)) = E_{(x,y) \sim P_{\text{data}}}
\left(- \sum_{t=1}^{m} \log P_{\theta}(y_t | y_{<t}, x)\right), \quad (1)$$

where $y_{<t}$ 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 an 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.

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

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:

$$L_{\text{fine-tune}}(P_{\text{target-data}}; \theta) + (\text{mix-ratio} \cdot \text{mix-decay}^{\text{epoch-id}}) \cdot L_{\text{pre-train}}(P_{\text{pretrain-data}}; \theta)$$  \hspace{1cm} (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 $\theta_{\text{pre}}$ (Kirkpatrick et al., 2016; Wiese et al., 2017). We denote it as WD($\theta_{\text{pre}}$) and formulate it as follows:

$$L_{\text{fine-tune}}(P_{\text{target-data}}; \theta) + \lambda \cdot \|\theta - \theta_{\text{pre}}\|^2_2$$  \hspace{1cm} (3)$$

In our experiments, we tune $\lambda$ 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 WD($\theta_{\text{pre}}$) with $\lambda = 0.1$. We observe that 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. 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.

\footnote{\url{http://commoncrawl.org/2016/10/news-dataset-available}}
Comparing to standard fine-tuning, mix-review further gives solid improvement. The gain is due to the discussion that the model pre-trained by NS better utilizes contextual input (which is further verified has more than 7% relative improvement. As discussed in Section 3.1, this confirms our earlier comparing to the baseline models trained from scratch. Comparing to MASS, the NS pre-training 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.

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. 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(θ 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(θ 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.

<table>
<thead>
<tr>
<th>Model(Data-set)</th>
<th>PPL(normal)</th>
<th>PPL(word-shuffle)</th>
<th>PPL(word-drop)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS Pre-trained(CCNEWS)</td>
<td>17.33</td>
<td>36.56(+110.96%)</td>
<td>35.56(+105.19%)</td>
</tr>
<tr>
<td>Baseline(Dailydialogue)</td>
<td>24.83</td>
<td>27.87(+12.2%)</td>
<td>31.87(+28.3%)</td>
</tr>
<tr>
<td>MASS+finetune(Dailydialogue)</td>
<td>12.78</td>
<td>15.85(+24.0%)</td>
<td>18.13(+41.8%)</td>
</tr>
<tr>
<td>NS+finetune(Dailydialogue)</td>
<td>11.54</td>
<td>16.30(+41.2%)</td>
<td>19.01(+64.7%)</td>
</tr>
<tr>
<td>NS+WD(θ pre)(Dailydialogue)</td>
<td>11.19</td>
<td>14.16(+26.5%)</td>
<td>16.37(+46.2%)</td>
</tr>
<tr>
<td>NS+Mix-review(Dailydialogue)</td>
<td>11.07</td>
<td>17.81(+60.8%)</td>
<td>23.05(+108.2%)</td>
</tr>
</tbody>
</table>

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 utilization of context, which explains its superior performance (in Section 5.1).
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dialogue Triggers</th>
<th>News Triggers</th>
<th>Dialogue Triggers</th>
<th>News Triggers</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS Pre-trained</td>
<td>0.245/0.089</td>
<td>0.347/0.153</td>
<td>0.245/0.089</td>
<td>0.347/0.153</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.124/0.007</td>
<td>0.101/0.004</td>
<td>0.032/0.0003</td>
<td>0.046/0.002</td>
</tr>
<tr>
<td>NS+finetune</td>
<td>0.162/0.047</td>
<td>0.158/0.046</td>
<td>0.187/0.052</td>
<td>0.170/0.044</td>
</tr>
<tr>
<td>NS+WD(θ_pre)</td>
<td>0.226/0.080</td>
<td>0.235/0.085</td>
<td>0.203/0.070</td>
<td>0.204/0.060</td>
</tr>
<tr>
<td>NS+Mix-review</td>
<td>0.261/0.108</td>
<td>0.322/0.135</td>
<td>0.223/0.079</td>
<td>0.341/0.151</td>
</tr>
</tbody>
</table>

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, 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.

---

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 closest 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.](https://parl.ai/)

\(^3\)https://parl.ai/

---

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

<table>
<thead>
<tr>
<th>Dailydialogue</th>
<th>Knowledge</th>
<th>Fluency</th>
<th>Consistency</th>
<th>Engagingness</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS+finetune</td>
<td>2.829 ± 0.061</td>
<td>5.359 ± 0.067</td>
<td>4.288 ± 0.067</td>
<td>3.848 ± 0.051</td>
</tr>
<tr>
<td>NS+WD(θ_{pre})</td>
<td>3.189 ± 0.067</td>
<td>5.071 ± 0.059</td>
<td>4.005 ± 0.064</td>
<td>4.184 ± 0.059</td>
</tr>
<tr>
<td>NS+Mix-review</td>
<td>3.401 ± 0.051</td>
<td>5.692 ± 0.060</td>
<td>4.755 ± 0.068</td>
<td>4.272 ± 0.056</td>
</tr>
</tbody>
</table>
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)$^\text{pre}$: Mix-review could be too “aggressive” and does not put enough attention on the target task. On the other hand, WD(θ)$^\text{pre}$ 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 knowledgeable 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


A BEAM-SEARCH VS. TOP-K SAMPLING

<table>
<thead>
<tr>
<th>Data-set</th>
<th>Beam Search</th>
<th>Top-30 Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Entropy</td>
<td>Max-ratio</td>
</tr>
<tr>
<td>Dailydialogue</td>
<td>7.44 8.49</td>
<td>1.7% 1.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Switchboard</td>
<td>4.96 5.54</td>
<td>34.9% 27.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cornell</td>
<td>6.10 6.56</td>
<td>10.2% 9.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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).
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 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 (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.

4http://compprag.christopherpotts.net/swda.html
5https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dials_Dialogs_Corpus.html
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).
Table 12: The detailed rating scores from AMT.

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

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.

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

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dialogue Triggers</th>
<th>News Triggers</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS Pre-trained</td>
<td>0.245/0.089</td>
<td>0.347/0.153</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.081/0.003</td>
<td>0.088/0.003</td>
</tr>
<tr>
<td>NS+finetune</td>
<td>0.207/0.071</td>
<td>0.207/0.063</td>
</tr>
<tr>
<td>NS+WD(θpre)</td>
<td>0.285/0.114</td>
<td>0.202/0.072</td>
</tr>
<tr>
<td>NS+Mix-review</td>
<td>0.396/0.190</td>
<td>0.212/0.065</td>
</tr>
</tbody>
</table>

Under review as a conference paper at ICLR 2020
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 $\text{WD}(\theta_{\text{pre}})$ could be not very effective for regularizing the model’s behavior.

Figure 3: Parameter-space UMAP projection of checkpoints from different training processes.