Garbage In, Flowers Out: Noisy Training Data Help Generative Models at Test Time

Alberto Testoni*
Università di Trento, Trento

Raffaella Bernardi**
Università di Trento, Rovereto

Despite important progress, conversational systems often generate dialogues that sound unnatural to humans. We conjecture that the reason lies in the different training and testing conditions: agents are trained in a controlled "lab" setting but tested in the "wild". During training, they learn to utter a sentence given the ground-truth dialogue history generated by human annotators. On the other hand, during testing, the agents must interact with each other, and hence deal with noisy data. We propose to fill this gap between the training and testing environments by training the model with mixed batches containing both samples of human and machine-generated dialogues. We assess the validity of the proposed method on GuessWhat?!, a visual referential game. We show that our method improves the linguistic quality of the generated dialogues, and it leads to higher accuracy of the guessing task; simple perturbations of the ground-truth dialogue history that mimic machine-generated data do not account for a similar improvement. Finally, we run a human evaluation experiment on a sample of machine-machine dialogues to complement the quantitative analysis. This experiment shows that also human annotators successfully exploit dialogues generated by a model trained with mixed batches to solve the task. Hence, the mixed-batch training does not cause a language drift. Moreover, we find that the new training regime allows human annotators to be significantly more confident when selecting the target object, showing that the generated dialogues are informative.

1. Introduction

Recent years have witnessed important progress in developing conversational agents, thanks to the introduction of the encoder-decoder framework (Sutskever, Vinyals, and Le 2014). This framework, in fact, allows learning directly from raw data for both natural language understanding and generation tasks. Promising results were obtained both for chit-chat (Vinyals and Le 2015) and task-oriented dialogues (Lewis et al. 2017). The framework has been further extended to develop agents that can communicate about a visual content using natural language (de Vries et al. 2017; Mostafazadeh et al. 2017; Das et al. 2017a). It is not easy to evaluate the performance of dialogue systems, but one crucial aspect is the quality of the generated dialogue. These systems must in fact produce a dialogue that sounds natural to humans in order to be employed in real-world scenarios. Although there is not a general agreement on what makes a machine-generated text sound natural, some features can be easily identified: for instance, natural language respects syntactic rules and semantic constraints, it is coherent, it contains

* Dept. of Information Engineering and Computer Science - DISI, Via Sommarive 9, 38123 Povo (TN) E-mail: alberto.testoni@unitn.it
** Center for Mind/Brain Sciences - CIMeC and Dept. of Information Engineering and Computer Science - DISI, Corso Bettini 31, 38068 Rovereto (TN) E-mail: raffaella.bernardi@unitn.it

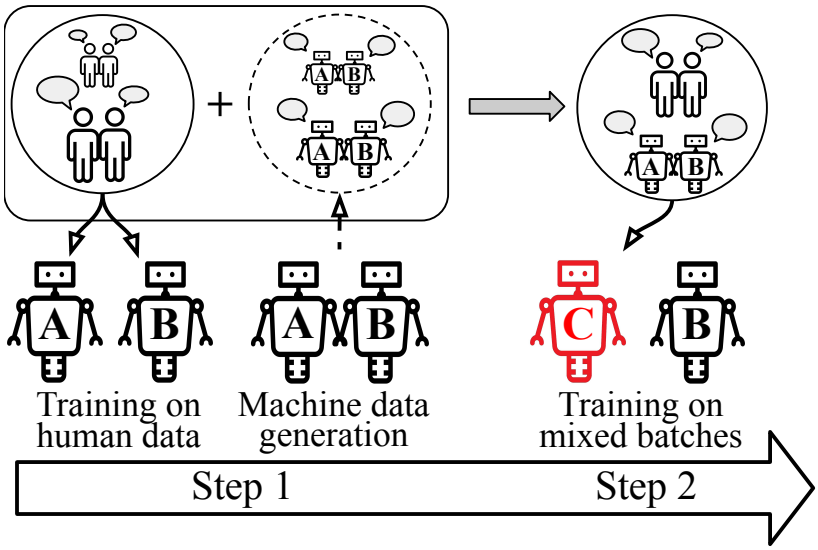

Figure 1
The two-step training method of the C Bot: two Bots, A and B, are trained independently to reproduce human dialogues; then they play together to generate new dialogues (step 1). In step 2 Bot C is trained on mixed batches of human and machine-generated data (by A and B in step 1).

words with different frequency distribution that, crucially, are informative for the conveyed message, and it does not have repetitions, both at token and sentence level.

Unfortunately, even state-of-the-art systems often generate dialogues that sound unnatural to humans, in particular due to the large number of repetitions appearing in the generated output. We conjecture that part of the problem is due to the training paradigm adopted by most of the systems. In the Supervised Learning training paradigm, the utterances generated by the models during training are used only to compute a Log Likelihood loss function with the gold-standard human dialogues and they are then thrown away. In a multi-turn dialogue setting, for instance, the follow-up utterance is always generated starting from the human dialogue, and not from the previously generated output. This procedure resembles a controlled "laboratory setting", where the agents are always exposed to "clean" human data at training time. Crucially, when tested, the agents are instead left alone "in the wild", without any human supervision. They have to "survive" in a new environment by exploiting the skills learned in the controlled lab setting and by interacting with each other.

Agents trained in a Reinforcement Learning fashion (Sutton and Barto 1998), on the other hand, are trained "in the wild" by maximizing a reward function based on the task success of the agent, at the cost of a significant increase in computational complexity. Agents trained according to this paradigm generate many repetitions and the quality of the dialogue degrades. This issue is mildly solved by the Cooperative Learning training (Shekhar et al. 2019b), but still, several repetitions occur in the dialogues, making them sound unnatural.

In this paper, we propose a simple but effective method to adjust the training environment so that it becomes more similar to the testing one (see Figure 1). In particular, we propose to replace part of the human training data with dialogues generated by conversational agents talking to each other (*Step 1* in Figure 1); these dialogues are

"noisy", since they may contain repetitions, a limited vocabulary etc. We then propose to train a new instance of the same conversational agent on this new training set ("mixed batches", *Step 2*). The model is now trained "out of the lab" since the data it is exposed to are less controlled and they get the model used to live in an environment more similar to the one it will encounter during testing.

We assessed the validity of the proposed method on a referential visual dialogue game, GuessWhat?! (de Vries et al. 2017). We found that the model trained according to our method outperforms the one trained only on human data with respect both to the accuracy in the guessing game and the linguistic quality of the generated dialogues, according to some surface-level linguistic metrics. In particular, the number of games with repeated questions drops significantly.

On a surface level, generated dialogues are mainly characterized by a high number of repetitions and a certain frequency of wrong answers coming from the Oracle agent, as we can see from the mixed batches sample dialogues illustrated in Figure 2. We checked the effect of these two features in isolation on ground-truth human dialogues by manually injecting them into the training set containing only human dialogues. We found that neither altering ground-truth answers nor injecting repetitions lead to an improvement at decoding time on the test set. We thus conclude that there exist more fine-grained features in the structure of machine-generated dialogues that lead to the improvement reported in this paper.[1]

2. Related Work

The need of going beyond the task success metric has been highlighted in Shekhar et al. (2019b), where the authors compare the quality of dialogues generated by their model against other state-of-the-art questioner models using some linguistic metrics. One striking feature of the dialogues generated by these models is the large number of dialogues containing repeated questions, while the ground-truth dialogues used to train the model (collected with human annotators) do not contain repetitions. In Shekhar et al. (2019a) the authors enrich the model proposed in Shekhar et al. (2019b) with a module that decides when the agent has gathered enough information and is ready to guess the target object. This approach is effective in reducing repetitions but, crucially, the task accuracy of the game decreases.

Murahari et al. (2019) propose a Questioner model for the GuessWhich task (Das et al. 2017b) that specifically aims to improve the diversity of generated dialogues by adding a new loss function during training: the authors propose a simple auxiliary loss that penalizes similar dialogue state embeddings in consecutive turns. Although this technique reduces the number of repeated questions compared to the baseline model, there is still a large number of repetitions in the output. Compared to these methods, our method does not require designing ad-hoc loss functions or plugging additional modules in the network.

The problem of generating repetitions affects not only dialogue systems, but instead it is a general property of current decoding strategies. Holtzman et al. (2020) found that decoding strategies that optimize for an output with high probability, such as the widely used beam and greedy search decoding, lead to a linguistic output that is

1 This paper is an extended version of Testoni and Bernardi (2020). We extended the previous work by investigating the role of specific features of machine-generated dialogues (Experiment 2 - Section 6.2) and by running a human evaluation experiment (Experiment 3 - Section 6.3).

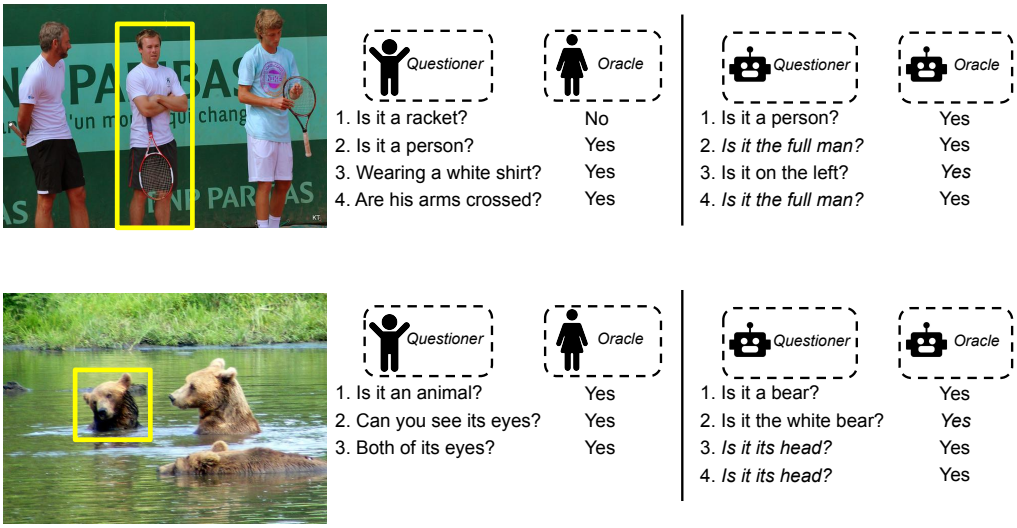

Figure 2
GuessWhat?! sample dialogues. Mixed batches consist of human-generated (left side) and machine-generated dialogues (right side, generated by GDSE-SL as in Shekhar et al., 2019). The yellow box highlights the target entity that the Questioner has to guess by asking binary questions to the Oracle. Both humans and conversational agents have to guess the target object only at the end of the dialogue. Note that the machine-generated dialogues on the right contain repetitions on the Questioner side, and wrong answers coming from the Oracle (both in *italic*).

generally banal, incoherent, and highly repetitive. Although language models generally assign high probabilities to well-formed text, the highest scores for longer texts are often repetitive and incoherent. To address this issue, the authors in Holtzman et al. (2020) propose a new decoding strategy that shows promising results, Nucleus Sampling. Starting from a probability distribution over all candidate tokens in the vocabulary, this technique samples the next token from the set of candidates defined as the top-p subset of the cumulative probability mass. Recently, Testoni and Bernardi (2021b) propose a beam-search re-ranking strategy to promote the generation of more effective questions throughout the dialogue. In this paper, we focus on the effect of different training sets using the same decoding strategy.

In a concurrent work, Suglia et al. (2021) propose a new training paradigm called Self-play via Iterated Experience Learning (SPIEL), in which the Questioner agent learns the task from games previously generated by other instances of the same Questioner architecture. The authors investigate to what extent the representations learned while playing guessing games can be transferred to other downstream tasks, such as VQA (Antol et al. 2015). Our approach is computationally simpler than the SPIEL technique: we train the same model in a Supervised Learning fashion using different training sets and we focus on the same task for all our experiments and training procedure.

3. Task and Models

Task. The GuessWhat?! game (de Vries et al. 2017) is a cooperative two-player game based on a referential communication task where two players collaborate to identify a

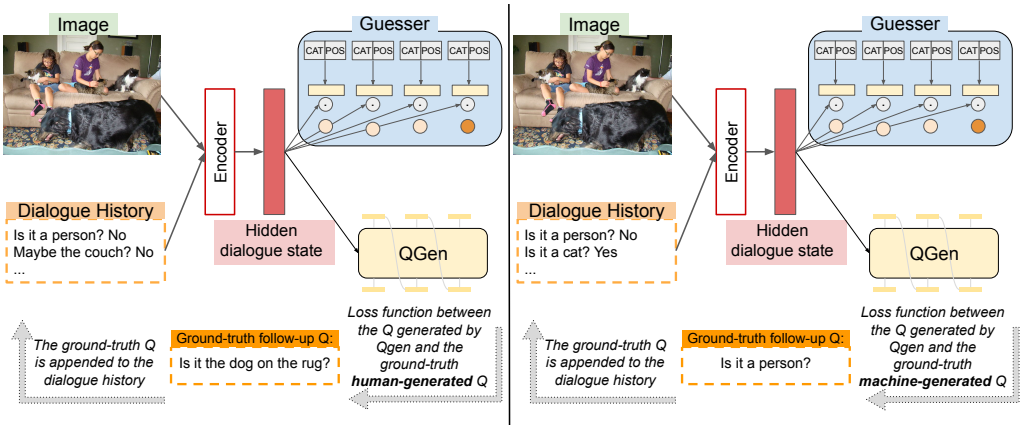

Figure 3
The image illustrates the follow-up question generation task investigated in this paper. Given an image and a dialogue history, the task consists in generating a follow-up question to identify the target in the image. During training, the loss function is computed by comparing the output of the QGen module with the ground-truth follow-up question: this question can be human-generated from the original GuessWhat?! training set (left), or machine-generated, i.e., generated by other instances of Questioner and Oracle agents interacting one with the other (right), as a result of *Step 1* illustrated in Image 1).

referent object in an image. This setting has been extensively used in human-human collaborative dialogue (Clark 1996; Yule 2013). GuessWhat?! is an asymmetric game involving two human participants who see a real-world image containing multiple objects from the MSCOCO dataset (Lin et al. 2014). One of the participants (the Oracle) is secretly assigned a target object within the image and the other participant (the Questioner agent) has to guess the target by asking binary (Yes/No) questions to the Oracle.

Models. We use the Visually-Grounded State Encoder (GDSE) model of Shekhar et al. (2019b), i.e. a Questioner agent for the GuessWhat?! game. We consider the version of GDSE trained in a supervised learning fashion (GDSE-SL). This model uses a visually grounded dialogue state that takes the visual features of the input image and each question-answer pair in the dialogue history to create a shared representation used both for generating a follow-up question (QGen module) and guessing the target object (Guesser module) in a multi-task learning scenario. More specifically, the visual features are extracted with a ResNet-152 network (He et al. 2016) and the dialogue history is encoded with an LSTM network (Hochreiter and Schmidhuber 1997). The full dialogue history consists of the concatenation of the individual dialogue turns, and each turn is composed of a question-answer pair. Since QGen faces a harder task and thus requires more training iterations, the authors made the learning schedule task-dependent. They called this setup *modulo-n* training, where *n* specifies after how many epochs of QGen training the Guesser component is updated together with QGen. The QGen component is optimized with the Log Likelihood of the training dialogues (as illustrated in Figure 3), and the Guesser computes a score for each candidate object by performing the dot product between visually grounded dialogue state and each object representation. As standard practice, the dialogues generated by the QGen are used only to compute the loss function, and the Guesser is trained by receiving human dialogues. At test time,

instead, the model generates a fixed number of questions (5 questions in our work) and the answers are obtained with the baseline Oracle agent presented in de Vries et al. (2017). Please refer to Shekhar et al. (2019b) for any additional detail on the model architecture and the training paradigm.

4. Metrics

Similarly to Testoni and Bernardi (2021a), the first metric we considered is the simple task accuracy (ACC) of the Questioner agent in guessing the target object among the candidates. We use four metrics to evaluate the quality of the generated dialogues.

- Games with repeated questions (GRQ), which measures the percentage of games with at least one repeated question verbatim.
- Mutual Overlap (MO), which represents the average of the BLEU-4 score obtained by comparing each question with the other questions within the same dialogue.
- Global Recall (GR), which measures the overall percentage of learnable words (i.e. words in the vocabulary) that the models recall (use) while generating new dialogues.
- Novel Questions (NQ), computed as the average number of questions in a generated dialogue that were not seen during training (compared via string matching).

The Mutual Overlap and Novel Questions metrics are taken from Murahari et al., (2019), while the Global Recall metric is taken from van Miltenburg et al., (2018). We acknowledge that these metrics do not capture the whole spectrum of the features that make a machine-generated text sound natural. However, GRQ, MO, and GR capture the most common weakness of neural generative models, namely the presence of repetitions (both token-level and sentence-level and reduced lexical variety). Therefore, we believe that these metrics represent a good proxy of the quality of the generated dialogues for the purpose of our study. The last metric, NQ, shed light on whether models simply learn to replicate human utterances or whether they are able to generate questions compositionally/creatively.

Finally, we complete the analysis with a human-based experiment by asking subjects to: (a) rate the grammaticality of the machine-generated dialogues, (b) select the target they describe, and (c) state their confidence when selecting the target. If humans find the dialogues grammatically sound and they are able to select the correct target, there is no language drift in place; if humans are highly confident of their choice, the dialogues are highly informative.

5. Datasets

We are interested in studying how modifying part of the human data in the training set affects the linguistic output and the model's accuracy on the GuessWhat?! game at test time. More specifically, we aim at building a training set in which part of the dialogues collected with human annotators are replaced with dialogues generated by the GDSE-SL questioner model while playing with the baseline Oracle model on the same games being replaced, i.e., the final training set size is always the same. In this way, we build

Table 1
Statistics of training sets built with different proportions of human machine-generated dialogues. Human data (100-0) vs. Mixed Batches (75-25, 50-50) vs. Fully Generated data (0-100). Voc size: size of the vocabulary used. GRQ: % games with at least one repeated question verbatim. MO: Mutual Overlap. Refer to Section 4 for additional details on the metrics.

% Human Dialogues	% Generated Dialogues	Generated Dial. Length	Voc size	MO↓	GRQ↓
100	0	variable	10469	0.05	0
75	25	fixed	4642	0.07	2.9
75	25	variable	4646	0.07	2.6
50	50	fixed	4391	0.08	5.4
50	50	variable	4396	0.07	4.7
0	100	fixed	2586	0.10	10.4
0	100	variable	2680	0.10	10.6

a training set containing dialogues that are more similar to the ones the model will generate at test time while playing with the Oracle.

Human data. The GuessWhat?! dataset was collected via Amazon Mechanical Turk by de Vries et al. (2017). The task is described in detail in Section 3. The training set contains about 108K dialogues and the validation and test set 23K each. Dialogues contain on average 5.2 turns, i.e, question-answer pairs exchanged between a Questioner and an Oracle. The images used in GuessWhat?! are taken from the MS-COCO dataset (Lin et al. 2014). Each image contains at least three and at most twenty objects. More than ten thousand people in total participated in the dataset collection procedure. Human annotators were instructed to ask only binary yes/no questions to the Oracle, and they could stop asking questions at any time, so the length of the dialogues is not fixed. Humans used a vocabulary of 17657 words to play GuessWhat?!: 10469 of these words appear at least three times, and thus make up the vocabulary given to the models. For our experiments, we considered only those games in which humans succeeded in identifying the target object in less than 20 dialogue turns, following a common practice from previous work.

Mixed Batches. We let the GDSE-SL model play with the baseline Oracle on the same games of the human training dataset. This produces automatically generated data for the whole training set. The model uses less than 3000 words out of a vocabulary of more than 10000 words. We built new training sets according to two criteria: the proportion of human and machine-generated data (50%-50% or 75%-25% respectively) and the length of the generated dialogue. Either we always keep a fixed dialogue length (5 turns, as the average length in the dataset) or we take the same number of turns that the human Questioner used while playing the game we are replacing.

Table 1 reports some statics of different training sets. Human dialogues have a very low mutual overlap and a much larger vocabulary than both the generated (100-0 setting, where the first number refers to the percentage of human-generated dialogues and the second one to the machine-generated percentage) and mixed batches datasets (50-50, 75-25). Looking at the number of games with at least one repeated question in the training set (GRQ column in Table 1), it can be observed that human annotators never produce dialogues with repetitions. The 75-25 dataset configuration contains less

than 3% of dialogues with repeated questions and this percentage rises to around 5% for the 50-50 configuration and to around 10% for the machine-generated dialogues (0-100). Looking at the vocabulary size, the human dataset (100-0) contains around ten thousand unique words, the mixed batches datasets (50-50, 75-25) around 4500 words, and the generated dialogues (0-100) approximately 2500 words.

Table 2
Experiment 1. Test Set 5Q setting. GDSE-SL results on several training sets. At test time, the model generates 5 questions and then it guesses. Length "fixed": 5-turns dialogues. Length "variable": same turns human annotators used for that game. ACC: accuracy. GRQ: % games with at least one repeated question. MO: Mutual Overlap. NQ: Novel Questions. GR: Global Recall. ↑: higher is better. ↓: lower is better. The difference in accuracy between the 100-0 setting and the other ones are all statistically significant, except for the one marked with *.

% Human Dialogues	% Machine Dialogues	Generated Dial. Length	ACC↑	GRQ↓	MO↓	NQ↑	GR↑
100	0	variable	46.3	36.8	0.27	0.53	20.6
75	25	fixed	47.9	24.0	0.20	0.43	20.2
75	25	variable	47.5	26.6	0.21	0.41	19.4
50	50	fixed	48.1	22.5	0.18	0.37	21.2
50	50	variable	47.0*	21.0	0.18	0.42	21.1

6. Experiments and Results

In the following, we report the results of three experiments aiming to understand how the training conditions affect accuracy and dialogue quality.

6.1 Experiment 1

In this first experiment, we evaluate models against the metrics described in Section 4 and compare the questioners whose training data differ with respect to the proportion of machine and human-generated dialogues.

6.1.1 Implementation Details

In order to obtain a dataset of machine-generated data (*Step 1* in Figure 1), we trained the GDSE-SL model for 100 epochs as described in Shekhar et al. (2019b). At the end of the training, we used GDSE to play the game with the Oracle on the whole training set, saving all the dialogues. We generate these dialogues with the model trained for all the 100 epochs since it generates fewer repetitions, although it is not the best-performing on the validation set. The dialogues generated by GDSE while playing with the Oracle are noisy: they may contain duplicated questions, wrong answers, etc. See Figure 2 for an example of human and machine-generated dialogues for the same game. We design different training sets as described in Section 5 and train the GDSE-SL model on these datasets (*Step 2* in Figure 1). We scrutinize the effect of training on different sets using the metrics described in Section 4 by letting the model generate new dialogues on the test set while playing with the Oracle.

6.1.2 Results

To have an overall comparison of the various settings, we computed the harmonic mean of all the metrics described above – after normalization and inverting the 'lower is better' metrics. The best balance is obtained with the 50-50 variable-length configuration. Below we zoom into the details of this coarse evaluation so as to highlight how the behaviour of the model changes across the settings.

Table 2 reports the results of the GDSE model trained on different training sets. By looking at the results on the test set, we can see how even a small number of machine-generated dialogues affects the generation phase at test time, when the model generates 5-turns dialogues and, at the end of the game, it guesses the target object. First of all, it can be noticed that the accuracy of GDSE-SL trained on the new datasets outperforms the one trained on the original training set: in particular, the accuracy of GDSE trained on 50% human dialogues and 50% 5-turns generated dialogues is almost 2% higher (in absolute terms) than the model trained only on human dialogues. The model seems to benefit from being exposed to noisy data at training time to better perform in the guessing game using the dialogues generated by the model itself while playing with the Oracle. We computed the McNemar's significance test comparing the accuracy of each setting with the 100-0 one: all the differences are statistically significant, with a significance level of 0.05, with the only exception of the 50-50 variable-length training.

The linguistic analysis of the dialogues generated on the test set reveals that the models trained on "mixed" batches produce better dialogues according to the metrics described in Section 4. In particular, considering the best-performing model on the test set, the percentage of games with repeated questions drops by 14.3% in absolute terms and the mutual overlap score by 0.09. The percentage of vocabulary used (global recall), on the other hand, remains stable. Interestingly, the only metric that seems to suffer from the model being trained on mixed datasets is the number of novel questions: being trained on noisy data does not seem to improve the "creativity" of the model, measured as the ability to generate new questions compared to ones seen at training time. The impact of the training data on the decoder's creativity and hence on the acquisition of human-like language generation ability deserves further investigation. However, it is out of the scope of our main research question, which instead focuses on understanding the quality of the generated dialogues rather than the creativity of the questioner.

Overall, our results show an interesting phenomenon: replacing part of the GuessWhat?! training set with machine-generated noisy dialogues, and training the GDSE-SL questioner model on this new dataset, is found to improve both the accuracy of the guessing game and the linguistic quality of the generated dialogues.

6.2 Experiment 2

Human dialogues and machine-generated ones have two easy-to-spot differences: the latter contain repetitions and the answers by the Oracle are not always correct (as discussed above and illustrated by Figure 2), while this is not the case for human dialogues. In this experiment, we aim to exclude the possibility that the results we obtained above are due to these two factors in isolation.

6.2.1 Implementational Details

In the following, we investigate more in detail the effect of the two surface-level features of machine-generated dialogues that are easier to inject in the training data: repetitions and wrong answers coming from the Oracle. We check whether one of these two

features alone accounts for the performance improvement described in the previous Section. To isolate their effect, we took the GuessWhat?! training set containing only ground-truth human-generated dialogues and we randomly swap the answers in the dialogue history or, alternatively, repeat one random question in the dialogue history. Considering that: the accuracy of the Oracle agent in answering questions is around 80%, and that the datasets used in the previous experiment contain on average 2.6% to 5.4% of games with repeated questions, we inject repetitions/wrong answers in the training set with a frequency of 5% or 20%.

Table 3
Experiment 2. Test-set results of GDSE-SL using training sets containing ground-truth dialogues with swapped answers or repetitions of questions in the dialogue history. The results are compared with Experiment 1 (Mixed Batches). ACC: accuracy. GRQ: % games with at least one repeated question. MO: Mutual Overlap. NQ: Novel Questions. GR: Global Recall. ↑: higher is better. ↓: lower is better.

		ACC ↑	**GRQ ↓**	**MO ↓**	**NQ ↑**	**GR ↑**
Mixed Batches Configuration	100 - 0	46.3	36.8	0.27	0.53	20.6
	50 - 50	48.1	22.5	0.18	0.37	21.2
% Swapped Answers	5	42.8	40.6	0.30	0.45	19.1
	20	32.1	48.7	0.34	0.43	17.2
GRQ (Training Set)	5	46.3	43.4	0.32	0.49	18.5
	20	44.0	65.2	0.43	0.37	18.4

6.2.2 Results

As we can see from Table 3 (top), swapping 5% or 20% of the answers in ground-truth human dialogue from the training set does not lead to the improvement we have presented in Section 6, but instead, it seriously degrades both the accuracy of the guessing task and the quality of the generated dialogues. More specifically, the accuracy drops by up to -14.2% compared to the model trained on the original GuessWhat?! dataset, and the percentage of games with at least one repeated question increases up to +11.9%.

Table 3 (bottom) shows the results of GDSE-SL when trained on a dataset of human dialogues containing a variable percentage of games with one duplicated question in the dialogue history. Compared to the dataset with swapped answers, in this case the drop in accuracy is very limited (up to -2.3%). On the other hand, the percentage of games with at least one repeated question at test time increases dramatically (up to +28.4% compared to the model trained on a dataset without repetitions), thus none of the parameters we consider improves.

Overall, these results demonstrate that the presence of repetitions or inconsistent answers alone in the training data do not account for the high performance illustrated in Section 6 using a training set composed of both human and machine-generated dialogues. This means that there are other features in the dialogue structure of generated dialogues that, when mixed with the human ground-truth data, create a training environment that, being closer to the inference setting, leads to more natural and effective dialogues.

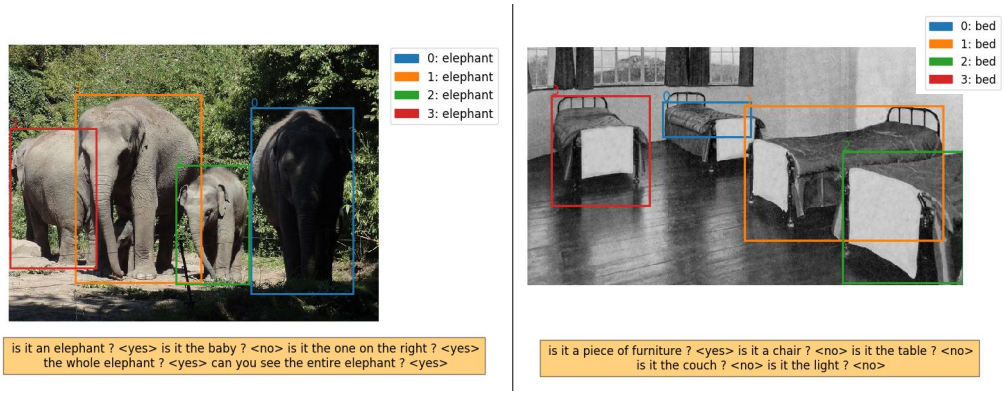

Figure 4
Human evaluation: subjects were given a machine-machine dialogue, an image, and the set of candidates among which they add to select the target and express their confidence. If the generated dialogue is informative the subject should be able to select the right target with high confidence, if the dialogue is not informative even if the selected target is correct the confidence should be low, as illustrated by the example on the left and on the right, respectively.

6.3 Experiment 3

The quantitative experiments above have shown that training the questioner with mixed-batches helps the model perform better at test time both in terms of task accuracy and in terms of the set of surface linguistics features we use as a proxy of dialogue quality. It remains open the question of whether the dialogues still sound natural, and hence would be suitable to interact with humans. To this end, we run a human-based evaluation and aim to exclude that our training protocol causes language drift. As claimed in Gauthier and Mordatch (2016), conversational agents must flexibly use human language to coordinate both with other agents and with humans to accomplish a goal.

6.3.1 Human Annotation Setup

We asked human annotators to read machine-generated dialogues and: 1. judge their grammaticality (using a scale from 1 to 5), 2. select the target object from a list of candidate objects, i.e., we asked human annotators to play the role of the Guesser module, 3. express their confidence on a scale from 1 to 5, their confidence in selecting the target object, where 5 means *'I am highly confident about my choice'* and 1 means *'My target selection was a random guess'*. We anticipate that all dialogues were evaluated as perfectly grammatical regardless of the training procedure, thus we do not discuss these results in the following.

We extracted 140 games from the GuessWhat validation set with fewer than 6 candidate objects in the image. For each game, we took the dialogue generated in the 100-0 and in the 50-50 paradigm. We recruited 4 annotators who were instructed about the GuessWhat task with some sample games[2] Each annotator was assigned to 70 dialogues (35 for the 100-0 setting, and 35 for the 50-50 setting, with no overlap between

2 Subjects provided written informed consent about the data collection.

the two sets). In this way, we obtained human annotation for 280 dialogues.[3] Figure 4 shows two examples of the images provided to the annotators.

Table 4
Results of the human evaluation experiment. *Human Accuracy* refers to the accuracy of human annotators in selecting the target object when reading machine-generated dialogues. *Confidence* refers to the average confidence of the annotators when selecting the target on a scale from 1(low confidence, random guess) to 5 (high confidence, steady selection).

		Human Accuracy ↑	Confidence ↑
Mixed Batches Configuration	100-0	66.43	3.51
	50-50	71.43	3.81

6.3.2 Results

Table 4 shows the results of our human evaluation experiment. They clearly demonstrate that the dialogues generated in the 50-50 setting do not show language drift. Although the accuracy of human annotators in identifying the target is higher in the 50-50 setting, the difference is not statistically significant according to a McNemar's statical test with a significance level of 0.05. Interestingly, human participants are significantly more confident when selecting the target in the mixed batches setting according to a paired t-test with a significance level of 0.05. These results show that the improvements of the mixed batches training regime highlighted in Table 2 do not come at the cost of generating dialogues that shift away from human ones. The dialogues results to be more informative also for humans who benefit from this new training paradigm.

7. Conclusion

Despite impressive progress in developing proficient conversational agents, current state-of-the-art systems produce dialogues that do not sound as natural as they should. In particular, they contain a high number of repetitions. To address this issue, methods presented so far in the literature implement new loss functions, or modify the models' architecture. When applied to referential guessing games, these techniques have the drawback of gaining little improvement, degrading the accuracy of the referential game, or producing incoherent dialogues.

Our work presents a simple but effective method to improve the linguistic output of conversational agents playing the GuessWhat?! game. Our idea starts from the observation that, in the Supervised Learning setting, models are trained in a "controlled" environment, i.e. using clean human-generated data while at test/inference time they are left "in the wild", and they have to solve the task by interacting one with the other without having access to ground-truth human-generated dialogues. Following this observation, in this paper we modify the GuessWhat?! training by replacing part of the dialogues produced by human annotators with machine-generated "noisy" dialogues, so that the training and testing conditions become more similar to each other. We show that a state-of-the-art model benefits from being trained on this new "mixed" dataset: being exposed to a small number of "noisy" dialogues at training time improves

3 We used the annotation tool available at this link to run the data collection: `https://www.makesense.ai/`.

the quality of the output without deteriorating its accuracy on the task. Our results show an absolute improvement in the accuracy of +1.8% and a drop in the number of dialogues containing duplicated questions of around -14%. We checked whether artificially injecting repetitions or swapping answers in ground-truth training data may lead to the same improvement, but we found this is not the case: more in-depth features of the machine-generated dialogues' structure make our "mixing" approach effective. Finally, to check that the generated dialogues were indeed effective also for human annotators and to rule out the hypothesis that language drift was behind the improvements described above, we run a human-based evaluation on a sample of the machine-machine dialogues. We found that human annotators not only successfully guess the target when reading dialogues generated with the new training regime, but they are also significantly more confident in their choice. We believe further work is required to check the effectiveness of our approach on other tasks/datasets, and to explore other kinds of perturbations on the input of generative neural dialogue systems.

References

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, Santiago, Chile, December 7-13. IEEE Computer Society.

Clark, Herbert H. 1996. *Using Language*. Cambridge University Press.

Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, Honolulu, Hawaii, USA, July.

Das, Abhishek, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. In *2017 IEEE International Conference on Computer Vision*, pages 2951–2960, Venice, Italy, October.

de Vries, Harm, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, Honolulu, Hawaii, USA, July.

Gauthier, Jon and Igor Mordatch. 2016. A paradigm for situated and goal-driven language learning. *arXiv preprint arXiv:1610.03585*.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, Las Vegas, NV, USA, June-July.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations (ICLR 2020)*, Addis Ababa, Ethiopia, April 26-30. OpenReview.net.

Lewis, Mike, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or No Deal? End-to-End learning for negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark, September.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of ECCV (European Conference on Computer Vision)*, pages 740–755, Zurich, Zwitzerland, September.

Mostafazadeh, Nasrin, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 462–472, Taipei, Taiwan, November.

Murahari, Vishvak, Prithvijit Chattopadhyay, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Improving generative visual dialog by answering diverse questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint*

Conference on Natural Language Processing, pages 1449–1454, Hong Kong, China, November.
Shekhar, Ravi, Alberto Testoni, Raquel Fernández, and Raffaella Bernardi. 2019a. Jointly Learning to See, Ask, Decide when to Stop, and then GuessWhat. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*, volume 2481 of *CEUR Workshop Proceedings*, Bari, Italy, November 13-15.
Shekhar, Ravi, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019b. Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2578–2587, Minneapolis, Mn, USA, June.
Suglia, Alessandro, Yonatan Bisk, Ioannis Konstas, Antonio Vergari, Emanuele Bastianelli, Andrea Vanzo, and Oliver Lemon. 2021. An empirical study on the generalization power of neural representations learned via visual guessing games. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2135–2144, Online, April. Association for Computational Linguistics.
Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3104–3112.
Sutton, Richard S. and Andrew G. Barto. 1998. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press.
Testoni, Alberto and Raffaella Bernardi. 2020. Overprotective training environments fall short at testing time: Let models contribute to their own training. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, volume 2769 of *CEUR Workshop Proceedings*, Bologna, Italy, March 1-3. CEUR-WS.org.
Testoni, Alberto and Raffaella Bernardi. 2021a. The interplay of task success and dialogue quality: An in-depth evaluation in task-oriented visual dialogues. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2071–2082, Online, April. Association for Computational Linguistics.
Testoni, Alberto and Raffaella Bernardi. 2021b. Looking for confirmations: An effective and human-like visual dialogue strategy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9330–9338, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
van Miltenburg, Emiel, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA, August.
Vinyals, Oriol and Quoc V. Le. 2015. A neural conversational model. In *Proceedings of the ICML Deep Learning Workshop*, Lille, France, July.
Yule, George. 2013. *Referential communication tasks*. Routledge.