Dialog-based Language Learning

Jason Weston
Facebook AI Research,
New York.
jase@fb.com

Abstract

A long-term goal of machine learning research is to build an intelligent dialog agent. Most research in natural language understanding has focused on learning from fixed training sets of labeled data, with supervision either at the word level (tagging, parsing tasks) or sentence level (question answering, machine translation). This kind of supervision is not realistic of how humans learn, where language is both learned by, and used for, communication. In this work, we study dialog-based language learning, where supervision is given naturally and implicitly in the response of the dialog partner during the conversation. We study this setup in two domains: the bAbI dataset of [23] and large-scale question answering from [3]. We evaluate a set of baseline learning strategies on these tasks, and show that a novel model incorporating predictive lookahead is a promising approach for learning from a teacher's response. In particular, a surprising result is that it can learn to answer questions correctly without any reward-based supervision at all.

1 Introduction

Many of machine learning's successes have come from supervised learning, which typically involves employing annotators to label large quantities of data per task. However, humans can learn by acting and learning from the consequences of (i.e, the feedback from) their actions. When humans act in dialogs (i.e., make speech utterances) the feedback is from other human's responses, which hence contain very rich information. This is perhaps most pronounced in a student/teacher scenario where the teacher provides positive feedback for successful communication and corrections for unsuccessful ones [8, 22]. However, in general any reply from a dialog partner, teacher or not, is likely to contain an informative training signal for learning how to use language in subsequent conversations.

In this paper we explore whether we can train machine learning models to learn from dialogs. The ultimate goal is to be able to develop an intelligent dialog agent that can learn *while conducting conversations*. To do that it needs to learn from feedback that is supplied as natural language. However, most machine learning tasks in the natural language processing literature are not of this form: they are either hand labeled at the word level (part of speech tagging, named entity recognition), segment (chunking) or sentence level (question answering) by labelers. Subsequently, learning algorithms have been developed to learn from that kind of supervision. We therefore need to develop evaluation datasets for the dialog-based language learning setting, as well as developing models and algorithms able to learn in such a regime.

The contribution of the present work is thus:

- We introduce a set of tasks that model natural feedback from a teacher and hence assess the feasibility of dialog-based language learning.
- We evaluate some baseline models on this data, comparing to standard supervised learning.
- We introduce a novel forward prediction model, whereby the learner tries to predict the teacher's replies to its actions, yielding promising results, even with no reward signal at all.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 Related Work

In human language learning the usefulness of social interaction and natural infant directed conversations is emphasized, see e.g. the review paper [6], although the usefulness of feedback for learning grammar is disputed [10]. Support for the usefulness of feedback is found however in second language learning [1] and learning by students [4, 8, 22].

In machine learning, one line of research has focused on supervised learning from dialogs using neural models [18, 3]. Question answering given either a database of knowledge [3] or short stories [23] can be considered as a simple case of dialog which is easy to evaluate. Those tasks typically do not consider feedback. There is work on the the use of feedback and dialog for learning, notably for collecting knowledge to answer questions [5, 14], the use of natural language instruction for learning symbolic rules [7] and the use of binary feedback (rewards) for learning parsers [2].

Another setting which uses feedback is the setting of reinforcement learning, see e.g. [16] for a summary of its use in dialog. However, those approaches often consider reward as the feedback model rather than exploiting the dialog feedback per se. Nevertheless, reinforcement learning ideas have been used to good effect for other tasks as well, such as understanding text adventure games [12], machine translation and summarization [15]. Recently, [11] also proposed a reward-based learning framework for learning how to learn.

Finally, forward prediction models, which we make use of in this work, have been used for learning eye tracking [17], controlling robot arms [9] and vehicles [21], and action-conditional video prediction in atari games [13]. We are not aware of their use thus far for dialog.

3 Dialog-Based Supervision Tasks

Dialog-based supervision comes in many forms. As far as we are aware it is a currently unsolved problem which type of learning strategy will work in which setting. In this section we therefore identify different modes of dialog-based supervision, and build a learning problem for each. The goal is to then evaluate learners on each type of supervision.

We thus begin by selecting two existing datasets: (i) the single supporting fact problem from the bAbI datasets [23] which consists of short stories from a simulated world followed by questions; and (ii) the MovieQA dataset [3] which is a large-scale dataset ($\sim 100k$ questions over $\sim 75k$ entities) based on questions with answers in the open movie database (OMDb). For each dataset we then consider ten modes of dialog-based supervision. The supervision modes are summarized in Fig. 1 using a snippet of the bAbI dataset as an example. The same setups are also used for MovieQA, some examples of which are given in Fig 2. We now describe the supervision setups.

Imitating an Expert Student In Task 1 the dialogs take place between a teacher and an expert student who gives semantically coherent answers. Hence, the task is for the learner to imitate that expert student, and become an expert themselves. For example, imagine the real-world scenario where a child observes their two parents talking to each other, it can learn but it is not actually taking part in the conversation. Note that our main goal in this paper is to examine how a non-expert can learn to improve its dialog skills while conversing. The rest of our tasks will hence concentrate on that goal. This task can be seen as a natural baseline for the rest of our tasks given the same input dialogs and questions.

Positive and Negative Feedback In Task 2, when the learner answers a question the teacher then replies with either positive or negative feedback. In our experiments the subsequent responses are variants of "No, that's incorrect" or "Yes, that's right". In the datasets we build there are 6 templates for positive feedback and 6 templates for negative feedback, e.g. "Sorry, that's not it.", "Wrong", etc. To separate the notion of positive from negative (otherwise the signal is just words with no notion that yes is better than no) we assume an additional *external reward* signal that is not part of the text. As shown in Fig. 1 Task 2, $(+)$ denotes positive reward external to the dialog (e.g. feedback provided by another medium, such as a nod of the head from the teacher). This is provided with every positive response. Note the difference in supervision compared to Task 1: there every answer is right and provides positive supervision. Here, only the answers the learner got correct have positive supervision. This could clearly be a problem when the learner is unskilled: it will supply incorrect answers and never (or hardly ever) receive positive responses.

Figure 1: Sample dialogs with differing supervision signals (tasks 1 to 10). In each case the same example story is given for simplicity (shown first). Black text is spoken by the teacher, red text denotes responses by the learner, blue text is provided by an expert student (which the learner can imitate), (+) denotes positive reward external to the dialog (e.g. feedback provided by another medium, such as a nod of the head from the teacher).

Story Context (same for all examples below)
Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.

Task 1: Imitating an Expert Student
Where is Mary? A:kitchen
Where is John? A:bathroom

Task 2: Positive and Negative Feedback
Where is Mary? A:playground
No, that's incorrect.
Where is John? A:bathroom
Yes, that's right! (+)

Task 3: Answers Supplied by Teacher
Where is Mary? A:bedroom
No, the answer is kitchen.
Where is John? A:bathroom
Correct! (+)

Task 4: Hints Supplied by Teacher
Where is Mary? A:bathroom
No, they are downstairs.
Where is John? A:kitchen
No, they are upstairs.

Task 5: Supporting Facts Supplied by Teacher
Where is Mary? A:kitchen
Yes, that's right! (+)
Where is John? A:hallway
No, because John moved to the bathroom.

Task 6: Partial Feedback
Where is Mary? A:kitchen
Yes, that's right!
Where is John? A:bathroom
Yes, that's correct! (+)

Task 7: No Feedback
Where is Mary? A:kitchen
Yes, that's right!
Where is John? A:bathroom
Yes, that's correct!

Task 8: Imitation and Feedback Mixture
Where is Mary? A:kitchen
Where is John? A:bathroom
That's right! (+)

Task 9: Asking For Corrections
Where is Mary? A:kitchen
Yes, that's right! (+)
Where is John? A:hallway
No, that's not right. A:Can you help me?
Bathroom.

Task 10: Asking For Supporting Facts
Where is Mary? A:kitchen
Yes, that's right! (+)
Where is John? A:hallway
No, that's not right. A:Can you help me?
A relevant fact is John moved to the bathroom.

Figure 2: Samples from the MovieQA dataset [3]. In our experiments we consider 10 different language learning setups as described in Figure 1 and Sec. 3. The examples given here are for tasks 2 and 3, questions are in black and answers in red, and (+) indicates receiving positive reward.

Task 2: Positive and Negative Feedback
What movies are about open source? Revolution OS
That's right! (+)
What movies did Darren McGavin star in? Carmen
Sorry, that's not it.
Who directed the film White Elephant? M. Curtiz
No, that is incorrect.

Task 3: Answers Supplied by Teacher
What films are about Hawaii? 50 First Dates
Correct! (+)
Who acted in Licence to Kill? Billy Madison
No, the answer is Timothy Dalton.
What genre is Saratoga Trunk in? Drama
Yes! (+)

Answers Supplied by Teacher In Task 3 the teacher gives positive and negative feedback as in Task 2, however when the learner's answer is incorrect, the teacher also responds with the correction. For example if "where is Mary?" is answered with the incorrect answer "bedroom" the teacher responds "No, the answer is kitchen"', see Fig. 1 Task 3. If the learner knows how to use this extra information, it effectively has as much supervision signal as with Task 1, and much more than for Task 2.

Hints Supplied by Teacher In Task 4, the corrections provided by the teacher do not provide the exact answer as in Task 3, but only a useful hint. This setting is meant to mimic the real life occurrence of being provided only partial information about what you did wrong. In our datasets

we do this by providing the *class* of the correct answer, e.g. "No, they are downstairs" if the answer should be kitchen, or "No, it is a director" for the question "Who directed Monsters, Inc.?" (using OMDB metadata). The supervision signal here is hence somewhere in between Task 2 and 3.

Supporting Facts Supplied by Teacher In Task 5, another way of providing partial supervision for an incorrect answer is explored. Here, the teacher gives a reason (explanation) why the answer is wrong by referring to a known fact that supports the true answer that the incorrect answer may contradict. For example "No, because John moved to the bathroom" for an incorrect answer to "Where is John?", see Fig. 1 Task 5. This is related to what is termed *strong supervision* in [23] where supporting facts and answers are given for question answering tasks.

Partial Feedback Task 6 considers the case where external rewards are only given some of (50% of) the time for correct answers, the setting is otherwise identical to Task 3. This attempts to mimic the realistic situation of some learning being more closely supervised (a teacher rewarding you for getting some answers right) whereas other dialogs have less supervision (no external rewards). The task attempts to assess the impact of such partial supervision.

No Feedback In Task 7 external rewards are not given at all, only text, but is otherwise identical to Tasks 3 and 6. This task explores whether it is actually possible to learn how to answer at all in such a setting. We find in our experiments the answer is surprisingly yes, at least in some conditions.

Imitation and Feedback Mixture Task 8 combines Tasks 1 and 2. The goal is to see if a learner can learn successfully from both forms of supervision at once. This mimics a child both observing pairs of experts talking (Task 1) while also trying to talk (Task 2).

Asking For Corrections Another natural way of collecting supervision is for the learner to ask questions of the teacher about what it has done wrong. Task 9 tests one of the most simple instances, where asking "Can you help me?" when wrong obtains from the teacher the correct answer. This is thus related to the supervision in Task 3 except the learner must first ask for help in the dialog. This is potentially harder for a model as the relevant information is spread over a larger context.

Asking for Supporting Facts Finally, in Task 10, a second less direct form of supervision for the learner after asking for help is to receive a hint rather than the correct answer, such as "A relevant fact is John moved to the bathroom" when asking "Can you help me?", see Fig. 1 Task 10. This is thus related to the supervision in Task 5 except the learner must request help.

In our experiments we constructed the ten supervision tasks for the two datasets which are all available for download at `http://fb.ai/babi`. They were built in the following way: for each task we consider a fixed policy[1] for performing actions (answering questions) which gets questions correct with probability π_{acc} (i.e. the chance of getting the red text correct in Figs. 1 and 2). We thus can compare different learning algorithms for each task over different values of π_{acc} (0.5, 0.1 and 0.01). In all cases a training, validation and test set is provided. For the bAbI dataset this consists of 1000, 100 and 1000 questions respectively per task, and for movieQA there are $\sim 96k$, $\sim 10k$ and $\sim 10k$ respectively. MovieQA also includes a knowledge base (KB) of $\sim 85k$ facts from OMDB, the memory network model we employ uses inverted index retrieval based on the question to form relevant memories from this set, see [3] for more details. Note that because the policies are fixed the experiments in this paper are *not* in a reinforcement learning setting.

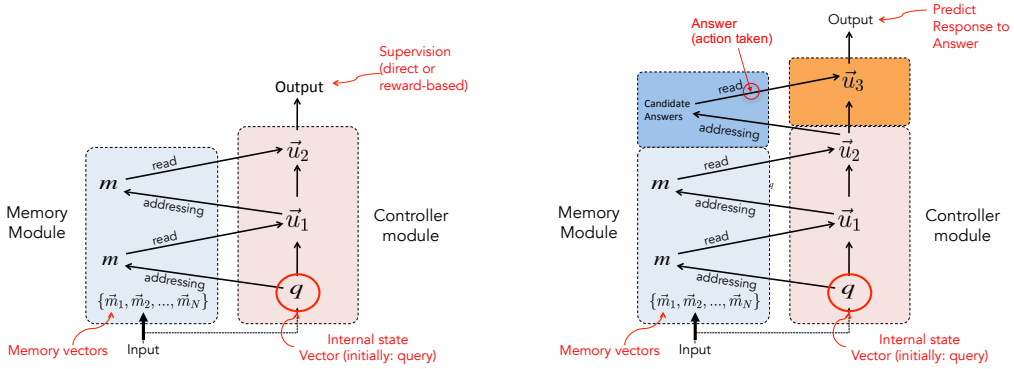
4 Learning Models

Our main goal is to explore training strategies that can execute dialog-based language learning. To this end we evaluate four possible strategies: imitation learning, reward-based imitation, forward prediction, and a combination of reward-based imitation and forward prediction. We will subsequently describe each in turn.

We test all of these approaches with the same model architecture: an end-to-end memory network (MemN2N) [20]. Memory networks are a recently introduced model that have been shown to do

[1] Since the policy is fixed and actually does not depend on the model being learnt, one could also think of it as coming from another agent (or the same agent in the past) which in either case is an imperfect expert.

Figure 3: Architectures for (reward-based) imitation and forward prediction.



(a) Model for (reward-based) imitation learning

(b) Model for forward prediction

well on a number of text understanding tasks, including question answering, dialog [3] and language modeling [20]. In particular, they outperform LSTMs and other baselines on the bAbI datasets [23] which we employ with dialog-based learning modifications in Sec. 3. They are hence a natural baseline model for us to use in order to explore differing modes of learning in our setup. In the following we will first review memory networks, detailing the explicit choices of architecture we made, and then show how they can be modified and applied to our setting of dialog-based language learning.

Memory Networks A high-level description of the memory network architecture we use is given in Fig. 3 (a). The input is the last utterance of the dialog, x, as well as a set of memories (context) $(c_1, \ldots, c_N)$ which can encode both short-term memory, e.g. recent previous utterances and replies, and long-term memories, e.g. facts that could be useful for answering questions. The context inputs c_i are converted into vectors m_i via embeddings and are stored in the memory. The goal is to produce an output $\hat{a}$ by processing the input x and using that to address and read from the memory, m, possibly multiple times, in order to form a coherent reply. In the figure the memory is read twice, which is termed multiple "hops" of attention.

In the first step, the input x is embedded using a matrix A of size $d \times V$ where d is the embedding dimension and V is the size of the vocabulary, giving $q = Ax$, where the input x is as a bag-of-words vector. Each memory c_i is embedded using the same matrix, giving $m_i = Ac_i$. The output of addressing and then reading from memory in the first hop is:

$$o_1 = \sum_i p_i^1 m_i, \quad p_i^1 = \text{Softmax}(q^\top m_i).$$

Here, the match between the input and the memories is computed by taking the inner product followed by a softmax, yielding p^1, giving a probability vector over the memories. The goal is to select memories relevant to the last utterance x, i.e. the most relevant have large values of p_i^1. The output memory representation o_1 is then constructed using the weighted sum of memories, i.e. weighted by p^1. The memory output is then added to the original input, $u_1 = R_1(o_1 + q)$, to form the new state of the controller, where R^1 is a $d \times d$ rotation matrix[2]. The attention over the memory can then be repeated using u_1 instead of q as the addressing vector, yielding:

$$o_2 = \sum_i p_i^2 m_i, \quad p_i^2 = \text{Softmax}(u_1^\top m_i),$$

The controller state is updated again with $u_2 = R_2(o_2 + u_1)$, where R_2 is another $d \times d$ matrix to be learnt. In a two-hop model the final output is then defined as:

$$\hat{a} = \text{Softmax}(u_2^\top A y_1, \ldots, u_2^\top A y_C) \tag{1}$$

[2]Optionally, different dictionaries can be used for inputs, memories and outputs instead of being shared.

where there are C candidate answers in y. In our experiments C is the set of actions that occur in the training set for the bAbI tasks, and for MovieQA it is the set of words retrieved from the KB.

Having described the basic architecture, we now detail the possible training strategies we can employ for our tasks.

Imitation Learning This approach involves simply imitating one of the speakers in observed dialogs, which is essentially a supervised learning objective[3]. This is the setting that most existing dialog learning, as well as question answer systems, employ for learning. Examples arrive as (x, c, a) triples, where a is (assumed to be) a good response to the last utterance x given context c. In our case, the whole memory network model defined above is trained using stochastic gradient descent by minimizing a standard cross-entropy loss between $\hat{a}$ and the label a.

Reward-based Imitation If some actions are poor choices, then one does not want to repeat them, that is we shouldn't treat them as a supervised objective. In our setting positive reward is only obtained immediately after (some of) the correct actions, or else is zero. A simple strategy is thus to only apply imitation learning on the rewarded actions. The rest of the actions are simply discarded from the training set. This strategy is derived naturally as the degenerate case one obtains by applying policy gradient [24] in our setting where the policy is fixed (see end of Sec. 3). In more complex settings (i.e. where actions that are made lead to long-term changes in the environment and delayed rewards) applying reinforcement learning algorithms would be necessary, e.g. one could still use policy gradient to train the MemN2N but applied to the model's own policy.

Forward Prediction An alternative method of training is to perform forward prediction: the aim is, given an utterance x from speaker 1 and an answer a by speaker 2 (i.e., the learner), to predict $\bar{x}$, the *response to the answer* from speaker 1. That is, in general to predict the changed state of the world after action a, which in this case involves the new utterance $\bar{x}$.

To learn from such data we propose the following modification to memory networks, also shown in Fig. 3 (b): essentially we chop off the final output from the original network of Fig. 3 (a) and replace it with some additional layers that compute the forward prediction. The first part of the network remains exactly the same and *only has access* to input x and context c, just as before. The computation up to $u_2 = R_2(o_2 + u_1)$ is thus exactly the same as before.

At this point we observe that the computation of the output in the original network, by scoring candidate answers in eq. (1) looks similar to the addressing of memory. Our key idea is thus to perform another "hop" of attention but over the candidate answers rather than the memories. Crucially, we also incorporate the information of which action (candidate) was actually selected in the dialog (i.e. which one is a). After this "hop", the resulting state of the controller is then used to do the forward prediction.

Concretely, we compute:

$$o_3 = \sum_i p_i^3 (A y_i + \beta^* [a = y_i]), \quad p_i^3 = \text{Softmax}(u_2^\top A y_i), \tag{2}$$

where β^* is a d-dimensional vector, that is also learnt, that represents in the output o_3 the action that was actually selected. After obtaining o_3, the forward prediction is then computed as:

$$\hat{x} = \text{Softmax}(u_3^\top A \bar{x}_1, \ldots, u_3^\top A \bar{x}_{\bar{C}})$$

where $u_3 = R_3(o_3 + u_2)$. That is, it computes the scores of the possible responses to the answer a over $\bar{C}$ possible candidates. The mechanism in eq. (2) gives the model a way to compare the most likely answers to x with the given answer a, which in terms of supervision we believe is critical. For example in question answering if the given answer a is incorrect and the model can assign high p_i to the correct answer then the output o_3 will contain a small amount of β^*; conversely, o_3 has a large amount of β^* if a is correct. Thus, o_3 informs the model of the likely response $\bar{x}$ from the teacher.

Training can then be performed using the cross-entropy loss between $\hat{x}$ and the label $\bar{x}$, similar to before. In the event of a large number of candidates $\bar{C}$ we subsample the negatives, always keeping $\bar{x}$ in the set. The set of answers y can also be similarly sampled, making the method highly scalable.

[3]Imitation learning algorithms are not always strictly supervised algorithms, they can also depend on the agent's actions. That is not the setting we use here, where the task is to imitate one of the speakers in a dialog.

Table 1: Test accuracy (%) on the Single Supporting Fact bAbI dataset for various supervision approachess (training with 1000 examples on each) and different policies π_{acc}. A task is successfully passed if $\geq 95\%$ accuracy is obtained (shown in blue).

	MemN2N *imitation learning*			MemN2N *reward-based imitation (RBI)*			MemN2N *forward prediction (FP)*			MemN2N *RBI + FP*		
Supervision Type π_{acc} =	0.5	0.1	0.01	0.5	0.1	0.01	0.5	0.1	0.01	0.5	0.1	0.01
1 - Imitating an Expert Student	100	100	100	100	100	100	23	30	29	99	99	100
2 - Positive and Negative Feedback	79	28	21	99	92	91	93	54	30	99	92	96
3 - Answers Supplied by Teacher	83	37	25	99	96	92	99	96	99	99	100	98
4 - Hints Supplied by Teacher	85	23	22	99	91	90	97	99	66	99	100	100
5 - Supporting Facts Supplied by Teacher	84	24	27	100	96	83	98	99	100	100	99	100
6 - Partial Feedback	90	22	22	98	81	59	100	100	99	99	100	99
7 - No Feedback	90	34	19	20	22	29	100	98	99	98	99	99
8 - Imitation + Feedback Mixture	90	89	82	99	98	98	28	64	67	99	98	97
9 - Asking For Corrections	85	30	22	99	89	83	23	15	21	95	90	84
10 - Asking For Supporting Facts	86	25	26	99	96	84	23	30	48	97	95	91
Number of completed tasks ($\geq 95\%$)	1	1	1	9	5	2	5	5	4	10	8	8

A major benefit of this particular architectural design for forward prediction is that after training with the forward prediction criterion, at test time one can "chop off" the top again of the model to retrieve the original memory network model of Fig. 3 (a). One can thus use it to predict answers $\hat{a}$ given only x and c. We can thus evaluate its performance directly for that goal as well.

Finally, and importantly, if the answer to the response $\bar{x}$ carries pertinent supervision information for choosing $\hat{a}$, as for example in many of the settings of Sec. 3 (and Fig. 1), then this will be backpropagated through the model. This is simply not the case in the imitation, reward-shaping [19] or reward-based imitation learning strategies which concentrate on the x, a pairs.

Reward-based Imitation + Forward Prediction As our reward-based imitation learning uses the architecture of Fig. 3 (a), and forward prediction uses the same architecture but with the additional layers of Fig 3 (b), we can learn jointly with both strategies. One simply shares the weights across the two networks, and performs gradient steps for both criteria, one of each type per action. The former makes use of the reward signal – which when available is a very useful signal – but fails to use potential supervision feedback in the subsequent utterances, as described above. It also effectively ignores dialogs carrying no reward. Forward prediction in contrast makes use of dialog-based feedback and can train without any reward. On the other hand not using rewards when available is a serious handicap. Hence, the mixture of both strategies is a potentially powerful combination.

Table 2: Test accuracy (%) on the MovieQA dataset dataset for various supervision approaches. Numbers in bold are the winners for that task and choice of π_{acc}.

	MemN2N *imitation learning*			MemN2N *reward-based imitation (RBI)*			MemN2N *forward prediction (FP)*			MemN2N *RBI + FP*		
Supervision Type π_{acc} =	0.5	0.1	0.01	0.5	0.1	0.01	0.5	0.1	0.01	0.5	0.1	0.01
1 - Imitating an Expert Student	**80**	**80**	**80**	80	80	80	24	23	24	77	77	77
2 - Positive and Negative Feedback	46	29	27	52	32	26	48	34	24	**68**	**53**	**34**
3 - Answers Supplied by Teacher	48	29	26	52	32	27	60	57	58	**69**	**65**	**62**
4 - Hints Supplied by Teacher	47	29	26	51	32	28	58	**58**	**42**	**70**	54	32
5 - Supporting Facts Supplied by Teacher	47	28	26	51	32	26	43	44	33	**66**	**53**	**40**
6 - Partial Feedback	48	29	27	49	32	24	60	58	58	**70**	**63**	**62**
7 - No Feedback	51	29	27	22	21	21	60	53	**58**	**61**	**56**	50
8 - Imitation + Feedback Mixture	60	50	47	63	53	51	46	31	23	**72**	**69**	**69**
9 - Asking For Corrections	48	29	27	52	34	26	67	**52**	**44**	**68**	**52**	39
10 - Asking For Supporting Facts	49	29	27	52	34	27	51	44	35	**69**	**53**	**36**
Mean Accuracy	52	36	34	52	38	34	52	45	40	**69**	**60**	**50**

5 Experiments

We conducted experiments on the datasets described in Section 3. As described before, for each task we consider a fixed policy for performing actions (answering questions) which gets questions correct with probability π_{acc}. We can thus compare the different training strategies described in Sec. 4 over each task for different values of π_{acc}. Hyperparameters for all methods are optimized on the validation sets. A summary of the results is reported in Table 1 for the bAbI dataset and Table 2 for MovieQA. We observed the following results:

- Imitation learning, ignoring rewards, is a poor learning strategy when imitating inaccurate answers, e.g. for $\pi_{acc} < 0.5$. For imitating an expert however (Task 1) it is hard to beat.
- Reward-based imitation (RBI) performs better when rewards are available, particularly in Table 1, but also degrades when they are too sparse e.g. for $\pi_{acc} = 0.01$.
- Forward prediction (FP) is more robust and has stable performance at different levels of π_{acc}. However as it only predicts answers implicitly and does not make use of rewards it is outperformed by RBI on several tasks, notably Tasks 1 and 8 (because it cannot do supervised learning) and Task 2 (because it does not take advantage of positive rewards).
- FP makes use of dialog feedback in Tasks 3-5 whereas RBI does not. This explains why FP does better with useful feedback (Tasks 3-5) than without (Task 2), whereas RBI cannot.
- Supplying full answers (Task 3) is more useful than hints (Task 4) but hints still help FP more than just yes/no answers without extra information (Task 2).
- When positive feedback is sometimes missing (Task 6) RBI suffers especially in Table 1. FP does not as it does not use this feedback.
- One of the most surprising results of our experiments is that FP performs well overall, given that it does not use feedback, which we will attempt to explain subsequently. This is particularly evident on Task 7 (no feedback) where RBI has no hope of succeeding as it has no positive examples. FP on the other hand learns adequately.
- Tasks 9 and 10 are harder for FP as the question is not immediately before the feedback.
- Combining RBI and FP ameliorates the failings of each, yielding the best overall results.

One of the most interesting aspects of our results is that FP works at all without any rewards. In Task 2 it does not even "know" the difference between words like "yes" or "'correct" vs. words like "wrong" or "incorrect", so why should it tend to predict actions that lead to a response like "yes, that's right"? This is because there is a natural coherence to predicting true answers that leads to greater accuracy in forward prediction. That is, you cannot predict a "right" or "wrong" response from the teacher if you don't know what the right answer is. In our experiments our policies π_{acc} sample negative answers equally, which may make learning simpler. We thus conducted an experiment on Task 2 (positive and negative feedback) of the bAbI dataset with a much more biased policy: it is the same as $\pi_{acc} = 0.5$ except when the policy predicts incorrectly there is probability 0.5 of choosing a random guess as before, and 0.5 of choosing the fixed answer *bathroom*. In this case the FP method obtains 68% accuracy showing the method still works in this regime, although not as well as before.

6 Conclusion

We have presented a set of evaluation datasets and models for dialog-based language learning. The ultimate goal of this line of research is to move towards a learner capable of talking to humans, such that humans are able to effectively teach it during dialog. We believe the dialog-based language learning approach we described is a small step towards that goal.

This paper only studies some restricted types of feedback, namely positive feedback and corrections of various types. However, potentially any reply in a dialog can be seen as feedback, and should be useful for learning. It should be studied if forward prediction, and the other approaches we tried, work there too. Future work should also develop further evaluation methodologies to test how the models we presented here, and new ones, work in those settings, e.g. in more complex settings where actions that are made lead to long-term changes in the environment and delayed rewards, i.e. extending to the reinforcement learning setting, and to full language generation. Finally, dialog-based feedback could also be used as a medium to learn non-dialog based skills, e.g. natural language dialog for completing visual or physical tasks.

Acknowledgments

We thank Arthur Szlam, Y-Lan Boureau, Marc'Aurelio Ranzato, Ronan Collobert, Michael Auli, David Grangier, Alexander Miller, Sumit Chopra, Antoine Bordes and Leon Bottou for helpful discussions and feedback, and the Facebook AI Research team in general for supporting this work.

References

[1] M. A. Bassiri. Interactional feedback and the impact of attitude and motivation on noticing l2 form. *English Language and Literature Studies*, 1(2):61, 2011.

[2] J. Clarke, D. Goldwasser, M.-W. Chang, and D. Roth. Driving semantic parsing from the world's response. In *Proceedings of computational natural language learning*, 2010.

[3] J. Dodge, A. Gane, X. Zhang, A. Bordes, S. Chopra, A. Miller, A. Szlam, and J. Weston. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931*, 2015.

[4] R. Higgins, P. Hartley, and A. Skelton. The conscientious consumer: Reconsidering the role of assessment feedback in student learning. *Studies in higher education*, 27(1):53–64, 2002.

[5] B. Hixon, P. Clark, and H. Hajishirzi. Learning knowledge graphs for question answering through conversational dialog. In *ACL*, 2015.

[6] P. K. Kuhl. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11): 831–843, 2004.

[7] G. Kuhlmann, P. Stone, R. Mooney, and J. Shavlik. Guiding a reinforcement learner with natural language advice: Initial results in robocup soccer. In *AAAI-2004 workshop on supervisory control*, 2004.

[8] A. S. Latham. Learning through feedback. *Educational Leadership*, 54(8):86–87, 1997.

[9] I. Lenz, R. Knepper, and A. Saxena. Deepmpc: Learning deep latent features for model predictive control. In *Robotics Science and Systems (RSS)*, 2015.

[10] G. F. Marcus. Negative evidence in language acquisition. *Cognition*, 46(1):53–85, 1993.

[11] T. Mikolov, A. Joulin, and M. Baroni. A roadmap towards machine intelligence. *arXiv preprint arXiv:1511.08130*, 2015.

[12] K. Narasimhan, T. Kulkarni, and R. Barzilay. Language understanding for text-based games using deep reinforcement learning. *arXiv preprint arXiv:1506.08941*, 2015.

[13] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in Neural Information Processing Systems*, pages 2845–2853, 2015.

[14] A. Pappu and A. Rudnicky. Predicting tasks in goal-oriented spoken dialog systems using semantic knowledge bases. In *Proceedings of the SIGDIAL*, pages 242–250, 2013.

[15] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

[16] V. Rieser and O. Lemon. *Reinforcement learning for adaptive dialogue systems*. Springer Science & Business Media, 2011.

[17] J. Schmidhuber and R. Huber. Learning to generate artificial fovea trajectories for target detection. *International Journal of Neural Systems*, 2(01n02):125–134, 1991.

[18] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. *NAACL*, 2015.

[19] P.-H. Su, D. Vandyke, M. Gasic, N. Mrksic, T.-H. Wen, and S. Young. Reward shaping with recurrent neural networks for speeding up on-line policy learning in spoken dialogue systems. *arXiv preprint arXiv:1508.03391*, 2015.

[20] S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, pages 2431–2439, 2015.

[21] G. Wayne and L. Abbott. Hierarchical control using networks trained with higher-level forward models. *Neural computation*, 2014.

[22] M. G. Werts, M. Wolery, A. Holcombe, and D. L. Gast. Instructive feedback: Review of parameters and effects. *Journal of Behavioral Education*, 5(1):55–75, 1995.

[23] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. Towards ai-complete question answering: a set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[24] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.