Countering Language Drift with KL Regularization

Michael Noukhovitch[♡♣♠]*, **Samuel Lavoie**[♡♣], **Issam H. Laradji**[♠],
Douwe Kiela[△], **Florian Strub**[□], **Aaron Courville**[♡♣♢]
[♡] Mila [♣] Université de Montréal [♠] ServiceNow Research
[△]Huggingface [□]Deepmind [♢] CIFAR Fellow
michael.noukhovitch@umontreal.ca

Abstract

End-to-end interactive learning of dialogue systems has been all-but-abandoned in favour of other approaches using more labelled data, such as dialogue state tracking. A major issue of the approach is that using language models as speaker and listener can lead to "language drift." Models are trained only to optimize a task objective and so their intermediate language can drift from pretrained natural language to an un-natural communication protocol. We reproduce previous work on tackling this phenomena and find that baseline methods are not as bad as reported. Furthermore, we use a simple KL regularization with an EMA model to stabilize RL training and outperform previous methods. Finally, we investigate the issue of "language drift" and find that it focuses only on the sender. We argue that "receiver drift" is equally important and show strong results on this novel metric.

1 Introduction

Dialogue agents that can interpret language and take actions in an environment are a long-term challenge for NLP. In particular, the methodology of training dialogue agents can be quite challenging. The current dominant paradigm is dialogue state tracking whereby dialogue is a supervised learning problem that relies of large amounts of labelled dialogue utterances, states, and actions. But labelling is a fundamentally difficult task and approaches that use less labelled data are of great interest. A less popular approach to dialogue is end-to-end learning that relies on training two (or more) dialogue agents to play both sides of the dialogue and train together. This approach leverages pretrained models and trains on the dialogue task directly allowing models to learn task-specific language and actions. But by optimizing two dialogue agents on a task without any language supervision can lead to "language drift" [Lee et al., 2019]. First demonstrated by Lewis et al. [2017], when two language models are trained with self-play on an end-to-end negitation task using RL, their communication drifts from their pretrained natural language to an in-human communication protocol with unknown semantics. Language drifts because agents are provided a reward solely for completing the task, and there is no language-level supervision requiring them to maintain a human-understandable language.

This problem has also been a major challenge in the field of Emergent Communication (EC), which investigates how two neural networks can communicate to solve a task [Wagner et al., 2003, Lazaridou and Baroni, 2020]. Previous approaches have sought to use a language model, another modality to ground communication [Lee et al., 2019], multi-tasking [Lowe* et al., 2021], and iterated learning [Lu et al., 2020a]. This work proposes that previous accounts of language drift have over-stated the severity of the problem and we propose a simple and efficient method to better counter language drift. First, we demonstrate that the severity of language drift has been overstated by reproducing previous work and, without hyperparameter tuning, achieve much better performance using the baseline model. Though previous work on language drift has focused on LSTMs, we demonstrate the same effect

*Work performed while visiting ServiceNow Research

InterNLP Workshop at 36th Conference on Neural Information Processing Systems (InterNLP @NeurIPS 2022).

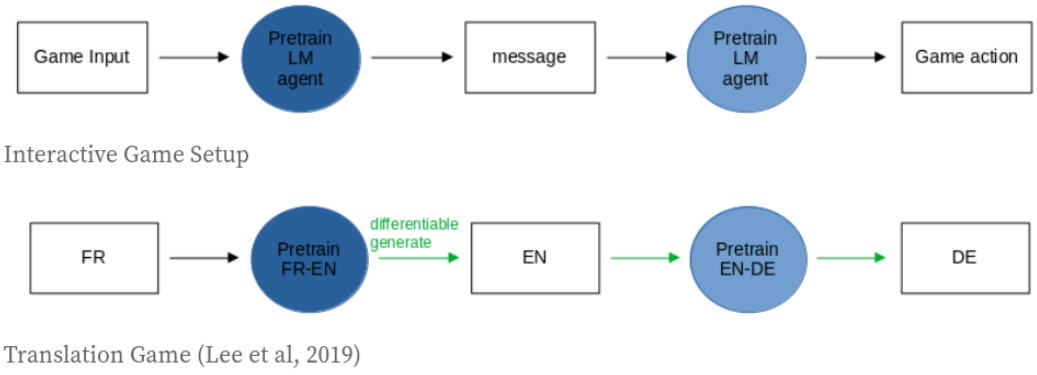


Figure 1: An interactive game with language compared to the Translation Game [Lee et al., 2019]

using Transformer models. Next, we examine previous performant methods to counter language drift and find that they are, to an extent, a tradeoff between maintaining the pretrained language and accomplishing the task. Inspired by recent practical and theoretical work, we use KL regularization between an online and EMA target network to counter language drift. Our method reduces language drift the most and yet can accomplish the task just as well as methods with more drift. Finally, we examine the practical origins of language drift and argue that previous drift metrics focused solely on the sender. We propose a novel drift metric for interactive language learning, receiver drift, and show why it may be more useful. We demonstrate strong results on the new metric that show promise for a resurgence of end-to-end methods for interactive language games.

2 Related Work

"Language Drift" in neural language models was first demonstrated by Lewis et al. [2017] by training two language models on a negotiation task and finding that the resulting syntax and semantics did not correspond to natural English. Since then, there has been a variety of work centered on countering language drift. Since it is difficult to measure the correctness of dialogue utterances, Lee et al. [2019] introduce the Translation Game as a benchmark task for language drift. Two models translate from French to English and English to German, respectively, and can measure accuracy using BLEU since the intermediate protocol is just English. They proposed a baseline that used a standard language model loss on the intermediate English representation to counter language drift, also later used by Steinert-Threlkeld et al. [2022]. They then outperformed the baseline by grounding the intermediate English in a separate modality, images. Lu et al. [2020a] proposed iterated learning and demonstrated improved performance on the same game. They later showed that multi-task learning on the original pretraining task (termed S2P by Lowe* et al. [2021]) was further effective at maintaining semantic coherence [Lu et al., 2020b]. Similarly, when finetuning GPT-3 [Brown et al., 2020] on human feedback using reinforcement learning, Ouyang et al. [2022] added the original pretraining objective to maintain linguistic consistency. On a different task, Lazaridou et al. [2020] argued that countering language drift at the level of semantics is best done by fixing the sender and learning to choose samples from it.

Language drift is also present when learning policies through latent language in RL [Andreas et al., 2018] and similarly, multi-tasking has been shown to be effective there [Jacob et al., 2021]. The pretrain-then-finetune setup with the goal of maintaining pretrained knowledge can be seen as a specific instance of continual learning and therefore language drift has links to catastrophic forgetting McCloskey and Cohen [1989]. There is a clear similarity between mitigation methods: rehearsal [Robins, 1995] or experience replay [Rolnick et al., 2019] is equivalent to multitasking with the pretraining objective [Lowe* et al., 2021] and weight-update regularization [Kirkpatrick et al., 2017] has similarities to the proposed KL regularization.

3 Setup

Translation Game We follow Lee et al. [2019] to set up the Translation Game. Two translation models, French to English (FR→EN) and English to German (EN→DE), are pretrained on IWSLT data [Cettolo et al., 2012]. Each model is a seq2seq [Sutskever et al., 2014] LSTM [Hochreiter and Schmidhuber, 1997] with attention [Bahdanau et al., 2015]. From the perspective of a sender-receiver game [Lewis, 1969] we consider FR→EN to be our sender and EN→DE to be our receiver. The models are given only paired French and German from Multi30k [Elliott et al., 2016, 2017] and learn to translate through an English pivot as shown in Figure 1. We measure success on the task as the overall FR→EN→DE BLEU score on the validation set. Since the models are not given English at finetune-time, we can measure the language drift as the drop in FR→EN BLEU on the validation set over training.

All experiments are run for 40k steps and, for fair comparison, all translation game models are initialized from the same pretrained models. We implement the translation game in the fairseq library [Ott et al., 2019] and run all experiments using 5 seeds where each run uses a single V100 GPU. All plots show the mean and standard error over seeds.

Baselines The EN→DE receiver is trained using cross-entropy between predicted and true DE. The main challenge is how to optimize the FR→EN sender by backpropogating the gradient through the discrete EN tokens. The most basic baseline is `frozen-sender`, we freeze the FR→EN sender and only update the EN→DE receiver. To train the sender, we must use a gradient estimator such as REINFORCE [Williams, 1992] as used by Lee et al. [2019] or Gumbel-Softmax [Jang et al., 2017, Maddison et al., 2017] as used by Lu et al. [2020a]. We choose `reinforce` with an exponentially moving baseline [2] and as with previous work, add a loss for entropy regularization.

To counter language drift, we implement previously introduced methods as baselines. We train an LSTM language model on IWSLT English text and use it to regularize the FR→EN sender. During the translation game, the `LM` baseline gets the negative log-likelihood of the sender's generated EN text under our trained language model and adds it to the sender's REINFORCE reward. We implement `multitask` learning i.e. S2P [Lowe* et al., 2021] by re-training the FR→EN sender on its pretraining data, IWSLT, as an auxiliary loss during the translation game. Finally, we implement Seeded Iterated Learning [`SIL`; Lu et al., 2020a] where agents alternate between the translation game and a form of knowledge distillation. A "teacher" sender and receiver train on the translation, then distill knowledge into a "student" sender and receiver respectively, finally the students are initialized as teachers for the next iteration.

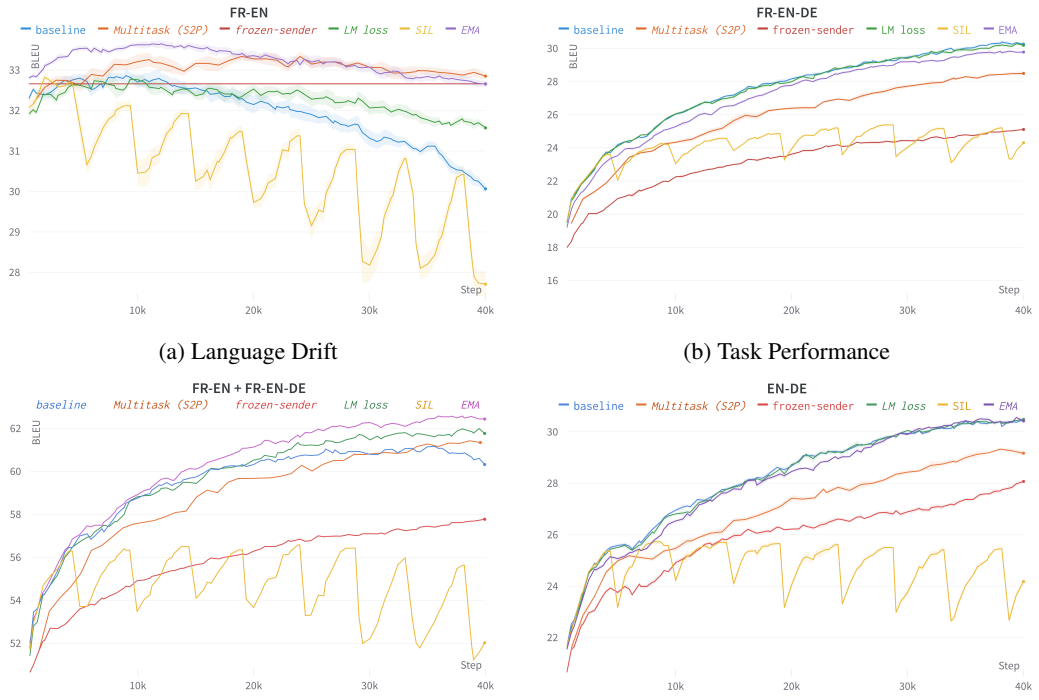
EMA KL Regularization We argue that language drift is a problem of noisy optimization. Recent theoretical [Geist et al., 2019, Vieillard et al., 2020b,a] and practical work in RL [Schulman et al., 2015, 2017] has demonstrated the efficacy of using a KL divergence between an online policy π_θ and target policy $\pi_{\bar{\theta}}$ to achieve stable training. We follow Chaabouni et al. [2022] and create a target policy by taking an exponential moving average (EMA) of our online policy's parameters during training $\bar{\theta} \leftarrow (1-\eta)\theta + \eta\bar{\theta}$ for some EMA parameter η. We create a target policy for the FR→EN sender and add an auxiliary KL loss between the online sender and target sender. Specifically, we generate EN text using the online FR→EN sender, and the auxiliary loss on the online network is the KL divergence between the logits of the online and target network. We refer to this method as `EMA`.

4 Experiments

First, we pretrain FR→EN and EN→DE models on IWSLT and reproduce IWSLT validation results from Lee et al. [2019], see Table 1 in Appendix A. We evaluate our pretrained models, zero-shot, on Multi30k for both FR→EN and FR→EN→DE and get better initial scores than Lee et al. [2019], Lu et al. [2020a] most likely due to better pretraining and preprocessing, see Appendix A.

Now, we interactively finetune our models on the translation game and plot our results in Figures 2a, 2b. As a sanity check, `frozen-sender` does not drift but also does not achieve a very high task performance. Most surprisingly, our `baseline` method, REINFORCE, does not drift much but still achieves relatively good task performance. This contrasts with previous work where baseline

[2] We also implemented Gumbel-Softmax and found similar, but slightly less performant results



(a) Language Drift

(b) Task Performance

(c) Combined Metrics

(d) Receiver Language Drift

Figure 2: Comparing methods on the Translation Game. We measure Language Drift (a) and Task Performance (b) over finetuning using FR→EN and FR→EN→DE BLEU respectively. Combined performance (c) is the sum FR→EN + FR→EN→DE BLEU and Receiver Language Drift (d), is our novel metric EN→DE BLEU.

REINFORCE or Gumbel-Softmax methods were reported to have high levels of drift. As well, our results with SIL [Lu et al., 2020a] are negative. We discuss both discrepancies further in Appenix B.

In line with previous results, we find that multi-tasking and an LM loss are both beneficial to reducing drift. But, as previously noted, these methods are auxiliary losses weighted by hyperparameters λ and the tuning of these hyperparameters is a direct tradeoff between drift and task performance. To demonstrate this, we choose a weight $\lambda = 0.01$ for our LM loss and find that it does not strongly impact task performance but neither does it strongly reduce drift. In contrast, we set our multitask weight $\lambda = 0.01$ and show it helps drift greatly but reduces task performance. Our KL regularization, `EMA`, manages to improve drift and also maintain high task performance. This is most evident when looking at the sum of our two metrics FR→EN + FR→EN→DE, as shown in Figure 2c, where `EMA` clearly maintains the highest sum. To confirm that these results are not artifacts of our model architecture, we implement the same setup using larger Transformer models [Vaswani et al., 2017]. We find similar results and demonstrate the efficacy of `EMA` in Appendix C.

Next, we reconsider our current metrics. Inspired by interactive dialogue, Lee et al. [2019] proposed to measure language drift using FR→EN BLEU. But this assumes that we would use the sender with humans after interactive training. Instead, it is just as reasonable to assume that our goal is to learn a receiver e.g. a flight booking system that takes natural language input and outputs a flight booking action. To this end, we should measure the receiver's language drift i.e. how much the receiver's understanding has drifted. In the translation game, we measure the DE BLEU score from inputting true EN text to our EN→DE receiver (as opposed to sender generated EN text in FR→EN→DE). We show results in Figure 2d and find, surprisingly, that all methods generally don't drift. Indeed, the REINFORCE `baseline` method performs as well as `EMA` and both outperform the `frozen-sender`. This implies that, in interactive training, the EN→DE receiver does not forget the true EN distribution while also learning the sender's drifted EN distribution.

5 Conclusion

We have demonstrated the efficacy of a simple KL regularization with an EMA model to counter language drift. Furthermore, we argue for a novel metric, receiver drift, and demonstrate that all methods perform well. With this in mind, we believe that interactive training with pretrained language models can be an effective training method, unencumbered by language drift.

References

J. Andreas, D. Klein, and S. Levine. Learning with Latent Language. In *NAACL*. arXiv, 2018. doi: 10.48550/arXiv.1711.00482. URL http://arxiv.org/abs/1711.00482. arXiv:1711.00482 [cs].

D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*. arXiv, 2015. URL http://arxiv.org/abs/1409.0473. arXiv:1409.0473 [cs, stat].

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In *Neural Information Processing Systems*. arXiv, July 2020. doi: 10.48550/arXiv.2005.14165. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

M. Cettolo, C. Girardi, and M. Federico. WIT3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy, May 2012. European Association for Machine Translation. URL https://aclanthology.org/2012.eamt-1.60.

R. Chaabouni, F. Strub, F. Altché, E. Tarassov, C. Tallec, E. Davoodi, K. W. Mathewson, O. Tieleman, A. Lazaridou, and B. Piot. Emergent Communication at Scale. In *International Conference on Learning Representations*, Mar. 2022. URL https://openreview.net/forum?id=AUGBfDIV9rL.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3210. URL http://aclweb.org/anthology/W16-3210.

D. Elliott, S. Frank, L. Barrault, F. Bougares, and L. Specia. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4718. URL https://aclanthology.org/W17-4718.

M. Geist, B. Scherrer, and O. Pietquin. A Theory of Regularized Markov Decision Processes. In *ICML*. arXiv, June 2019. doi: 10.48550/arXiv.1901.11275. URL http://arxiv.org/abs/1901.11275. arXiv:1901.11275 [cs, stat].

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, Nov. 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

A. P. Jacob, M. Lewis, and J. Andreas. Multitasking Inhibits Semantic Drift. In *NAACL*. arXiv, Apr. 2021. doi: 10.48550/arXiv.2104.07219. URL http://arxiv.org/abs/2104.07219. arXiv:2104.07219 [cs].

E. Jang, S. Gu, and B. Poole. Categorical Reparameterization with Gumbel-Softmax. In *ICLR*. arXiv, Aug. 2017. doi: 10.48550/arXiv.1611.01144. URL http://arxiv.org/abs/1611.01144. arXiv:1611.01144 [cs, stat].

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, Mar. 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/10.1073/pnas.1611835114. Publisher: Proceedings of the National Academy of Sciences.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-2045.

A. Lazaridou and M. Baroni. Emergent Multi-Agent Communication in the Deep Learning Era, July 2020.

A. Lazaridou, A. Potapenko, and O. Tieleman. Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning. In *ACL*. arXiv, May 2020. doi: 10.48550/arXiv.2005.07064. URL http://arxiv.org/abs/2005.07064. arXiv:2005.07064 [cs].

J. Lee, K. Cho, and D. Kiela. Countering Language Drift via Visual Grounding, Sept. 2019. URL http://arxiv.org/abs/1909.04499. arXiv:1909.04499 [cs].

D. Lewis. *Convention: A Philosophical Study*. Wiley-Blackwell, 1969.

M. Lewis, D. Yarats, Y. N. Dauphin, D. Parikh, and D. Batra. Deal or No Deal? End-to-End Learning for Negotiation Dialogues, June 2017. URL http://arxiv.org/abs/1706.05125. arXiv:1706.05125 [cs] version: 1.

R. Lowe*, A. Gupta*, J. Foerster, D. Kiela, and J. Pineau. On the interaction between supervision and self-play in emergent communication. In *International Conference on Learning Representations*, Sept. 2021. URL https://openreview.net/forum?id=rJxGLlBtwH.

Y. Lu, S. Singhal, F. Strub, O. Pietquin, and A. Courville. Countering Language Drift with Seeded Iterated Learning. In *ICML*. arXiv, Aug. 2020a. doi: 10.48550/arXiv.2003.12694. URL http://arxiv.org/abs/2003.12694. arXiv:2003.12694 [cs].

Y. Lu, S. Singhal, F. Strub, O. Pietquin, and A. Courville. Supervised Seeded Iterated Learning for Interactive Language Learning. In *EMNLP*. arXiv, Oct. 2020b. doi: 10.48550/arXiv.2010.02975. URL http://arxiv.org/abs/2010.02975. arXiv:2010.02975 [cs].

C. J. Maddison, A. Mnih, and Y. W. Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *ICLR*. arXiv, Mar. 2017. doi: 10.48550/arXiv.1611.00712. URL http://arxiv.org/abs/1611.00712. arXiv:1611.00712 [cs, stat].

M. McCloskey and N. J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In G. H. Bower, editor, *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, Jan. 1989. doi: 10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North*, pages 48–53, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-4009. URL http://aclweb.org/anthology/N19-4009.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, Mar. 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

M. Post. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319.

A. Robins. Catastrophic Forgetting, Rehearsal and Pseudorehearsal. *Connection Science*, 7 (2):123–146, June 1995. ISSN 0954-0091. doi: 10.1080/09540099550039318. URL https://doi.org/10.1080/09540099550039318. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/09540099550039318.

D. Rolnick, A. Ahuja, J. Schwarz, T. P. Lillicrap, and G. Wayne. Experience Replay for Continual Learning, Nov. 2019. URL http://arxiv.org/abs/1811.11682. arXiv:1811.11682 [cs, stat].

J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel. Trust Region Policy Optimization. In *ICML*. arXiv, 2015. doi: 10.48550/arXiv.1502.05477. URL http://arxiv.org/abs/1502.05477. arXiv:1502.05477 [cs].

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal Policy Optimization Algorithms, Aug. 2017. URL http://arxiv.org/abs/1707.06347. arXiv:1707.06347 [cs].

S. Steinert-Threlkeld, X. Zhou, Z. Liu, and C. M. Downey. Emergent Communication Fine-tuning (EC-FT) for Pretrained Language Models. In *Emergent Communication Workshop at ICLR 2022*, June 2022. URL https://openreview.net/forum?id=SUqrM7WR7W5.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to Sequence Learning with Neural Networks. In *Neural Information Processing Systems*. arXiv, Dec. 2014. URL http://arxiv.org/abs/1409.3215. arXiv:1409.3215 [cs].

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All you Need. In *Neural Information Processing Systems*, page 11, 2017.

N. Vieillard, T. Kozuno, B. Scherrer, O. Pietquin, R. Munos, and M. Geist. Leverage the Average: an Analysis of KL Regularization in RL. In *Neural Information Processing Systems*. arXiv, 2020a. doi: 10.48550/arXiv.2003.14089. URL http://arxiv.org/abs/2003.14089. arXiv:2003.14089 [cs, stat].

N. Vieillard, O. Pietquin, and M. Geist. Munchausen Reinforcement Learning. In *Neural Information Processing Systems*. arXiv, Nov. 2020b. doi: 10.48550/arXiv.2007.14430. URL http://arxiv.org/abs/2007.14430. arXiv:2007.14430 [cs, stat].

K. Wagner, J. A. Reggia, J. Uriagereka, and G. S. Wilkinson. Progress in the Simulation of Emergent Communication and Language. *Adaptive Behavior*, 11(1):37–69, Mar. 2003. ISSN 1059-7123, 1741-2633. doi: 10.1177/10597123030111003. URL http://journals.sagepub.com/doi/10.1177/10597123030111003.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, page 28, 1992.

A IWSLT Pretraining

We follow Lee et al. [2019] for preprocessing and pretraining, with the addition of using early stopping on the IWSLT validation set (tst2013) to choose our pretrained model.

	FR→EN	EN→DE
Lee et al. [2019]	34.1	22.0
Lu et al. [2020a]	32.2	20.2
Ours	38.5	23.2

Table 1: BLEU score of IWSLT-pretrained models on IWSLT 2013 validation set

B Translation Game Results

	Method	FR→EN	FR→EN→DE	Drift	Perf	Combined
Lee et al. [2019]	pretrained	27.2	16.3			
	REINFORCE	12.4 ± 0.7	24.5 ± 1.5	-14.8	+8.2	-6.6
	+ LM	23.6 ± 1.1	27.7 ± 0.4	-3.6	+11.4	+7.8
	+ LM + G	24.8 ± 0.4	28.1 ± 0.7	-2.4	+11.8	+9.4
Lu et al. [2020a]	pretrained	29.4	15.7			
	gumbel-softmax	14.5 ± 0.8	27.1 ± 0.1	-14.9	+11.4	-3.5
	SIL	29.4 ± 0.3	28.3 ± 0.2	0	+12.6	+12.6
Ours	pretrained	32.6	18.0			
	frozen-sender	32.6 ± 0	25.1 ± 0.1	0	+7.1	+7.1
	REINFORCE	30.0 ± 0.2	30.3 ± 0.2	-2.6	+12.3	+9.7
	LM $\lambda = 0.01$	31.5 ± 0.1	30.2 ± 0.2	-1.1	+12.2	+11.1
	Multitask $\lambda = 0.1$	32.9 ± 0.2	28.5 ± 0.3	+0.3	+10.5	+10.8
	SIL	27.7 ± 0.7	24.3 ± 0.1	-4.9	6.3	+1.4
	EMA $\lambda = 2$	32.6 ± 0.1	29.8 ± 0.1	0	+11.8	**+11.8**

Table 2: BLEU scores and pm standard deviation on the Multi30k Translation Game using IWSLT-pretrained LSTM models. Drift is the negative change in FR→EN BLEU from the pretrained model. Task performance is the positive change in FR→EN→DE BLEU from the pretrained models. Combined is the sum of drift and task performance. We show the pretrained, baseline, and best-performing model from previous work. Note that our results are not directly comparable to previous work because we evaluate using detokenized sacreBLEU [Post, 2018] whereas previous work wrote their own BLEU evaluation code and did not detokenize.

Our results for the baseline are notably better than Lee et al. [2019], Lu et al. [2020a]. The only difference between our code and theirs as far as we can tell is

1. we use an exponential moving average baseline for REINFORCE whereas [Lee et al., 2019] use an Actor-Critic method
2. Lu et al. [2020a] uses 0.1 gradient clipping and we do not use gradient clipping
3. in preprocessing Multi30k, we first tokenize [Koehn et al., 2007] then lowercase whereas previous works did the inverse.

A more reasonable explanation for the improvement in results is how we choose to evaluate. Previous work simply ran all methods for the same number of updates but this doesn't account for, even implicit, hyperparameter optimization. Previous methods show that the baseline's FR→EN→DE BLEU scores plateau and there is significant language drift in FR→EN without real improvements to the task score. We hypothesize that these extra training episodes only serve to increase the drift without measuring what we actually care about: performance gain for drift. Since the number of updates is arbitrary, we believe that early stopping on a reasonable metric is a better evaluation protocol.

To that end, we choose our combined metric to be the sum of the two metrics FR→EN→DE and FR→EN which we plot in Figure 2c. We choose learning rates so that all methods "peak" at around

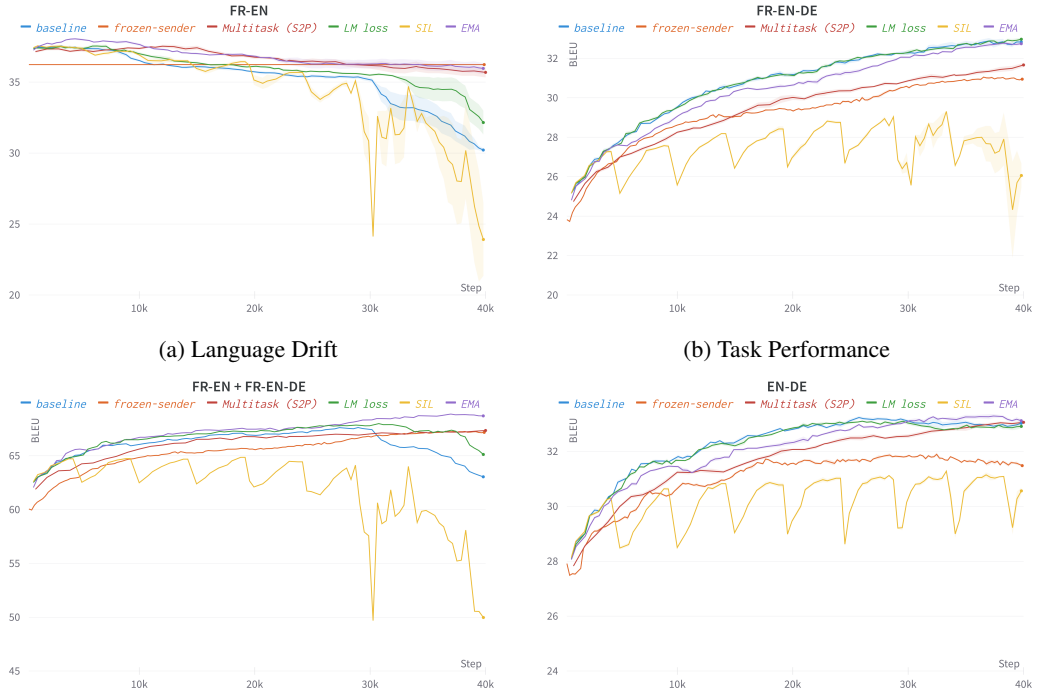
40k updates. Looking at the combined metrics, we see that our EMA method outperforms others by a reasonable margin i.e. methods that perform on FR→EN→DE drift more and methods that do as well on FR→EN do not reach as high a task performance.

We also note that our results with the SIL method of Lu et al. [2020a] are negative. We do not manage to gain any improvement in performance. We collaborated with the authors of Lu et al. [2020a] for many months but, in our setup, could not reproduce their results. At its core, we could not reproduce their fundamental results: a teacher sender can be outperformed by the student sender that it is distilling to.

C Translation Game with Transformers

We replicate the same experiment using transformer models to demonstrate that our method is not restricted to the LSTM architecture. We use a 6-layer Transformer-small architecture and follow the standard IWSLT translation setup of Ott et al. [2019]. We plot results in Figure 3 and find results similar to those with an LSTM in Figure 2.

Once again, the `frozen-sender` sanity check does not drift but neither does it perform as well. `Multitask` and `LM` are both tradeoffs between drift and performance. Results with `SIL` are still negative. And `EMA` is still the best-performing model, clearly visible when looking at the combined metrics in Figure 3c.



(a) Language Drift

(b) Task Performance

(c) Combined Metrics

(d) Receiver Language Drift

Figure 3: Comparing methods with a Transformer architecture on the Translation Game. We measure Language Drift (a) and Task Performance (b) over finetuning using FR→EN and FR→EN→DE BLEU respectively. Combined performance (c) is the sum FR→EN + FR→EN→DE BLEU and Receiver Language Drift (d), is our novel metric EN→DE BLEU.