Self-Regulated Interactive Sequence-to-Sequence Learning

Julia Kreutzer
Computational Linguistics
Heidelberg University
Germany
kreutzer@cl.uni-heidelberg.de

Stefan Riezler
Computational Linguistics & IWR
Heidelberg University
Germany
riezler@cl.uni-heidelberg.de

Abstract

Not all types of supervision signals are created equal: Different types of feedback have different costs and effects on learning. We show how self-regulation strategies that decide when to ask for which kind of feedback from a teacher (or from oneself) can be cast as a learning-to-learn problem leading to improved cost-aware sequence-to-sequence learning. In experiments on interactive neural machine translation, we find that the self-regulator discovers an ϵ-greedy strategy for the optimal cost-quality trade-off by mixing different feedback types including corrections, error markups, and self-supervision. Furthermore, we demonstrate its robustness under domain shift and identify it as a promising alternative to active learning.

1 Introduction

The concept of self-regulation has been studied in educational research (Hattie and Timperley, 2007; Hattie and Donoghue, 2016), psychology (Zimmerman and Schunk, 1989; Panadero, 2017), and psychiatry (Nigg, 2017), and was identified as central to successful learning. "Self-regulated students" can be characterized as "becoming like teachers", in that they have a repertoire of strategies to self-assess and self-manage their learning process, and they know when to seek help and which kind of help to seek. While there is a vast literature on machine learning approaches to meta-learning (Schmidhuber et al., 1996), learning-to-learn (Thrun and Pratt, 1998), or never-ending learning (Mitchell et al., 2015), the aspect of learning when to ask for which kind of feedback has so far been neglected in this field.

We propose a machine learning algorithm that uses self-regulation in order to balance the cost and effect of learning from different types of feedback. This is particularly relevant for human-in-the-loop machine learning, where human supervision is costly. The self-regulation module automatically learns which kind of feedback to apply when in training—full supervision by teacher demonstration or correction, weak supervision in the form of positive or negative rewards for student predictions, or a self-supervision signal generated by the student. Figure 1 illustrates this learning scenario. The learner, in our case a sequence-to-sequence (Seq2Seq) learner, aims to solve a certain task with the help of a human teacher. For every input it receives for training, it can ask the teacher for feedback to its own output, or supervise itself by training on its own output, or skip learning on the input example altogether. The self-regulator's policy for choosing feedback types is guided by their cost and by the performance gain achieved by learning from a particular type of feedback.

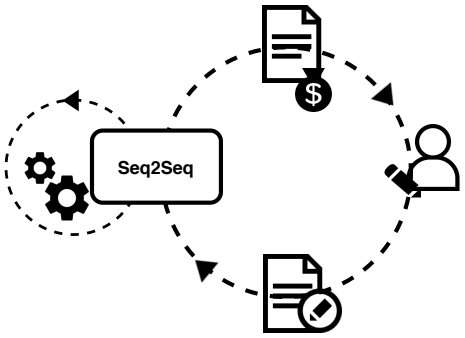


Figure 1: Human-in-the-loop self-regulated learning.

We apply the self-regulation algorithm to interactive machine translation where a neural machine translation (NMT) system functions as a student which receives feedback simulated from a human reference translation or supervises itself. The intended real-world application is a machine translation personalization scenario where the goal of the human translator is to teach the NMT system

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 303–315
Florence, Italy, July 28 - August 2, 2019.

to adapt to in-domain data with the best trade-off between feedback cost and performance gain. It can be transferred to other sequence-to-sequence learning tasks such as personalization of conversational AI systems for question-answering or geographical navigation.

Our analysis of different configurations of self-regulation yields the following insights: Perhaps unsurprisingly, the self-regulator learns to balance all types of feedback instead of relying only on the strongest or cheapest option. This is an advantage over active learning strategies that only consider the choice between no supervision and full supervision. Interestingly, though, we find that the self-regulator learns to trade off exploration and exploitation similar to a context-free ϵ-greedy strategy that optimizes ϵ for fastest learning progress. Lastly, we show that the learned regulator is robust in a cold-start transfer to new domains, and even shows improvements over fully supervised learning on domains such as literary books where reference translations provide less effective learning signals.

2 Related Work

The incorporation of a query's cost into reinforcement learning has been addressed, for example, in the framework of active reinforcement learning (Krueger et al., 2016). The central question in active reinforcement learning is to quantify the long-term value of reward information, however, assuming a fixed cost for each action and every round. Our framework is considerably more complicated by the changing costs for each feedback type on each round.

A similar motivation for the need of changing feedback in reinforcement learning with human feedback is given in MacGlashan et al. (2017). The goal of that work is to operationalize feedback schemes such as diminishing returns, differential feedback, or policy shaping. Human reinforcement learning with corrective feedback that can decrease or increase the action magnitude has been introduced in Celemin et al. (2019). However, none of these works are concerned with the costs that are incurred when eliciting rewards from humans, nor do they consider multiple feedback modes.

Our work is connected to active learning, for example, to approaches that use reinforcement learning to learn a policy for a dynamic active learning strategy (Fang et al., 2017), or to learn a curriculum to order noisy examples (Kumar et al., 2019), or to the approach of Liu et al. (2018) who use imitation learning to select batches of data to be labeled. However, the action space these approaches consider is restricted to the decision whether or not to select particular data and is designed for a fixed budget, neither do they incorporate feedback cost in their frameworks. As we will show, our self-regulation strategy outperforms active learning based on uncertainty sampling (Settles and Craven, 2008; Peris and Casacuberta, 2018) and our reinforcement learner is rewarded in such a way that it will produce the best system as early as possible.

Research that addresses the choice and the combination of different types of feedback is situated in the area between reinforcement and imitation learning (Ranzato et al., 2016; Cheng et al., 2018). Instead of learning how to mix different supervision signals, these approaches assume fixed schedules.

Further connections between our work on learning with multiple feedback types can be drawn to various extensions of reinforcement learning by multiple tasks (Jaderberg et al., 2017), multiple loss functions (Wun et al., 2018), or multiple policies (Smith et al., 2018).

Feedback in the form of corrections (Turchi et al., 2017), error markings (Domingo et al., 2017), or translation quality judgments (Lam et al., 2018) has been successfully integrated in simulation experiments into interactive-predictive machine translation. Again, these works do not consider automatic learning of a policy for the optimal choice of feedback.

3 Self-Regulated Interactive Learning

In this work, we focus on the aspect of self-regulated learning that concerns the ability to decide which type of feedback to query from a teacher (or oneself) for most efficient learning depending on the context. In our human-in-the-loop machine learning formulation, we focus on two contextual aspects that can be measured precisely: quality and cost. The self-regulation task is to optimally balance human effort and output quality.

We model self-regulation as an active reinforcement learning problem with dynamic costs, where in each state, i.e. upon receiving an input, the regulator has to choose an action, here a feedback

type, and pay a cost. The learner receives feedback of that type from the human to improve its prediction. Based on the effect of this learning update, the regulator's actions are reinforced or penalized, so that it improves its choice for future inputs.

In the following, we first compare training objectives for a Seq2Seq learner from various types of feedback (§3.1), then introduce the self-regulator module (§3.2), and finally combine both in the self-regulation algorithm (§3.3).

3.1 Seq2Seq Learning

Let $x = x_1 \dots x_S$ be a sequence of indices over a source vocabulary $\mathcal{V}_{\text{SRC}}$, and $y = y_1 \dots y_T$ a sequence of indices over a target vocabulary $\mathcal{V}_{\text{TRG}}$. The goal of sequence-to-sequence learning is to learn a function for mapping an input sequence x into an output sequences y. Specifically, for the example of machine translation, where y is a translation of x, the model, parametrized by a set of weights θ, learns to maximize $p_\theta(y \mid x)$. This quantity is further factorized into conditional probabilities over single tokens:

$$p_\theta(y \mid x) = \prod_{t=1}^{T} p_\theta(y_t \mid x; y_{<t}).$$

The distribution $p_\theta(y_t \mid x; y_{<t})$ is defined by the neural model's softmax-normalized output vector:

$$p_\theta(y_t \mid x; y_{<t}) = \texttt{softmax}(\text{NN}_\theta(x; y_{<t})).$$

There are various options for building the architecture of the neural model NN_θ, such as recurrent (Sutskever et al., 2014), convolutional (Gehring et al., 2017) or attentional (Vaswani et al., 2017) encoder-decoder architectures (or a mix thereof (Chen et al., 2018)). Regardless of their architecture, there are multiple ways of interactive learning that can be applied to neural Seq2Seq learners.

Learning from Corrections (FULL). Under full supervision, i.e., when the learner receives a fully corrected output y^* for an input x, cross-entropy minimization (equivalent to maximizing the likelihood of the data $\mathcal{D}$ under the current model) considers the following objective:

$$J^{\text{FULL}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,y^*)\in\mathcal{D}} -\log p_\theta(y^* \mid x).$$

The stochastic gradient of this objective is

$$g_\theta^{\text{FULL}}(x, y^*) = -\nabla_\theta \log p_\theta(y^* \mid x),$$

constituting an unbiased estimate of the gradient

$$\nabla_\theta J^{\text{FULL}} = \mathbb{E}_{(x,y^*)\sim\mathcal{D}} \left[g_\theta^{\text{FULL}}(x, y^*) \right].$$

A local minimum can be found by performing stochastic gradient descent on $g_\theta^{\text{FULL}}(x, y^*)$. This training objective is the standard in supervised learning when training with human-generated references or for online adaptation to post-edits (Turchi et al., 2017).

Learning from Error Markings (WEAK). Petrushkov et al. (2018) presented chunk-based binary feedback as a low-cost alternative to full corrections. In this scenario the human teacher marks the correct parts of the machine-generated output $\hat{y}$. As a consequence every token in the output receives a reward δ_t, either $\delta_t = 1$ if marked as correct, or $\delta_t = 0$ otherwise. The objective of the learner is to maximize the likelihood of the correct parts of the output, or equivalently, to minimize

$$J^{\text{WEAK}}(\theta) = \tfrac{1}{|\mathcal{D}|} \textstyle\sum_{(x,\hat{y})\in\mathcal{D}} \sum_{t=1}^{T} -\delta_t \log p_\theta(\hat{y}_t \mid x; \hat{y}_{<t})$$

where the stochastic gradient is

$$g_\theta^{\text{WEAK}}(x, \hat{y}) = -\sum_{t=1}^{T} \delta_t \cdot \nabla_\theta \log p_\theta(\hat{y}_t \mid x; y_{<t})$$

$$\nabla_\theta J^{\text{WEAK}} = \mathbb{E}_{(x,\hat{y})\sim\mathcal{D}} \left[g_\theta^{\text{WEAK}}(x, \hat{y}) \right].$$

The tokens $\hat{y}_t$ that receive $\delta_t = 1$ are part of the correct output y^*, so the model receives a hint how a corrected output should look like. Although the likelihood of the incorrect parts of the sequence does not weigh into the sum, they are contained in the context of the correct parts (in $y_{<t}$).

Self-Supervision (SELF). Instead of querying the teacher for feedback, the learner can also choose to learn from its own output, that is, to learn from self-supervision. The simplest option is to treat the learner's output as if it was correct, but that quickly leads to overconfidence and degeneration. Clark et al. (2018) proposed a cross-view training method: the learner's original prediction is used as a target for a weaker model that shares parameters with the original model. We adopt this strategy by first producing a target sequence $\hat{y}$ with beam search and then weaken the decoder through attention dropout with probability p_{att}. The objective is to minimize the negative likelihood of the original target under the weakened model

$$J^{\text{SELF}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,\hat{y})\in\mathcal{D}} -\log p_\theta^{p_{att}}(\hat{y} \mid x),$$

where the stochastic gradient is

$$g_\theta^{\text{SELF}}(x, \hat{y}) = -\nabla_\theta \log p_\theta^{p_{att}}(\hat{y} \mid x)$$
$$\nabla_\theta J^{\text{SELF}} = \mathbb{E}_{(x,\hat{y})\sim\mathcal{D}}\left[g_\theta^{\text{SELF}}(x, \hat{y})\right].$$

Combination. For self-regulated learning, we also consider a fourth option (NONE): the option to ignore the current input. Figure 2 summarizes the stochastic gradients for all cases.

$$g_\theta^s(x, y) = -\sum_{t=1}^{T} f_t \cdot \nabla_\theta \log p_\theta^{drop}(y_t \mid x_t; y_{<t}),$$
$$\text{with } y = \begin{cases} y^* & \text{if } s = \text{FULL} \\ \hat{y} & \text{otherwise,} \end{cases}$$
$$drop = \begin{cases} p_{\text{att}} & \text{if } s = \text{SELF} \\ 0 & \text{otherwise,} \end{cases}$$
$$\text{and } f_t = \begin{cases} 1 & \text{if } s \in \{\text{FULL}, \text{SELF}\} \\ \delta_t & \text{if } s = \text{WEAK} \\ 0 & \text{if } s = \text{NONE} \end{cases}$$

Figure 2: Stochastic gradients for the Seq2Seq learner in dependence of feedback type s.

In practice, Seq2Seq learning shows greater stability for mini-batch updates than online updates on single training samples. Mini-batch self-regulated learning can be achieved by accumulating stochastic gradients for a mini-batch of size $\mathcal{B}$ before updating θ with an average of these stochastic gradients, which we denote as $g_\theta^{s_{[1:\mathcal{B}]}}(x_{[1:\mathcal{B}]}, y_{[1:\mathcal{B}]}) = \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} g_\theta^{s_i}(x_i, y_i)$.

3.2 Learning to Self-Regulate

The regulator is another neural model q_ϕ that is optimized for the quality-cost trade-off of the Seq2Seq learner. Given an input x_i and the Seq2Seq's hypothesis $\hat{y}_i$, it chooses an action, here a supervision mode $s_i \sim q_\phi(s \mid x_i, \hat{y}_i)$. This choice of feedback determines the update of the Seq2Seq learner (Figure 2). The regulator is rewarded by the ratio between the cost c_i of obtaining the feedback s_i and the quality improvement $\Delta(\theta_i, \theta_{i-1})$ caused by updating the Seq2Seq learner with the feedback:

$$r(s_i, x_i, \theta_i) = \frac{\Delta(\theta_i, \theta_{i-1})}{c_i + \alpha}. \quad (1)$$

$\Delta(\theta_i, \theta_{i-1})$ is measured as the difference in validation score achieved before and after the learner's update (Fang et al., 2017), and c_i as the cost of user edits. Adding a small constant cost α to the actual feedback cost ensures numerical stability. This meta-parameter can be interpreted as representing a basic cost for model updates of any kind.

The objective for the regulator is to maximize the expected reward defined in Eq. 1:

$$J^{\text{META}}(\phi) = \mathbb{E}_{x\sim p(x), s\sim q_\phi(s|x,\hat{y})}\left[r(s, x, \theta)\right].$$

The full gradient of this objective is estimated by the stochastic gradient for sampled actions (Williams, 1992):

$$g_\phi^{\text{META}}(x, \hat{y}, s) = r \cdot \nabla_\phi \log q_\phi(s \mid x, \hat{y}). \quad (2)$$

Note that the reward contains the immediate improvement after one update of the Seq2Seq learner and not the overall performance in hindsight. This is an important distinction to classic expected reward objectives in reinforcement learning since it biases the regulator towards actions that have an immediate effect, which is desirable in the case of interaction with a human. However, since Seq2Seq learning requires updates and evaluations based on mini-batches, the regulator update also needs to be based on mini-batches of predictions, leading to the following specification of Eq. (2) for a mini-batch j:

$$g_\phi^{\text{META}}(x_{[1:\mathcal{B}]}, \hat{y}_{[1:\mathcal{B}]}, s_{[1:\mathcal{B}]}) \quad (3)$$
$$= \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} g_\phi^{\text{META}}(x_i, \hat{y}_i, s_i)$$
$$= \Delta(\theta_j, \theta_{j-1}) \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \frac{\nabla_\phi \log q_\phi(s_i \mid x_i, \hat{y}_i)}{c_i + \alpha}.$$

While mini-batch updates are required for stable Seq2Seq learning, they hinder the regulator from assigning credit for model improvement to individual elements within the mini-batch.

3.3 Algorithm

Algorithm 1 presents the proposed online learning algorithm with model updates cumulated over mini-batches. On arrival of a new input, the regulator predicts a feedback type in line 6. According to this prediction, the environment/user is asked for feedback for the Seq2Seq's prediction at cost c_i (line 7). The Seq2Seq model is updated on the

Algorithm 1 Self-Regulated Interactive Seq2Seq

Input: Initial Seq2Seq θ_0, regulator ϕ_0, $\mathcal{B}$

1: $j \leftarrow 0$
2: **while** inputs and human available **do**
3: $\quad j \leftarrow j + 1$
4: $\quad$ **for** $i \leftarrow 1$ to $\mathcal{B}$ **do**
5: $\quad\quad$ Observe input x_i, Seq2Seq output $\hat{y}_i$
6: $\quad\quad$ Choose feedback: $s_i \sim q_\phi(s \mid x_i, \hat{y}_i)$
7: $\quad\quad$ Obtain feedback f_i of type s_i at cost c_i
8: $\quad$ Update θ with $g_\theta^{s_{[1:\mathcal{B}]}}(x_{[1:\mathcal{B}]}, \hat{y}_{[1:\mathcal{B}]})$
9: $\quad$ Measure improvement $\Delta(\theta_j, \theta_{j-1})$
10: $\quad$ Update ϕ with $g_\phi^{\text{META}}(x_{[1:\mathcal{B}]}, \hat{y}_{[1:\mathcal{B}]}, s_{[1:\mathcal{B}]})$

basis of the feedback and mini-batch of stochastic gradients computed as summarized in Figure 2. In order to reinforce the regulator, the Seq2Seq model's improvement (line 9) is assessed, and the parameters of the regulator are updated (line 10, Eq. 3). Training ends when the data stream or the provision of feedback ends. The intermediate Seq2Seq evaluations can be re-used for model selection (early stopping). In practice, these evaluations can either be performed by validation on a held-out set (as in the simulation experiments below) or by human assessment.

Practical Considerations. The algorithm does not introduce any additional hyperparameters beyond standard learning rates, architecture design and mini-batch sizes that have to be tuned. As proposed in Petrushkov et al. (2018) or Clark et al. (2018), targets $\hat{y}$ are pre-generated offline with the initial θ_0, which we found crucial for the stability of the learning process. The evaluation step after the Seq2Seq update is an overhead that comes with meta-learning, incurring costs depending on the decoding algorithm and the evaluation strategy. However, Seq2Seq updates can be performed in mini-batches, and the improvement is assessed after a mini-batch of updates, as discussed above.

4 Experiments

The main research questions to be answered in our experiments are:

1. Which strategies does the regulator develop?
2. How well does a trained regulator transfer across domains?
3. How do these strategies compare against (active) learning from a single feedback type?

We perform experiments for interactive NMT, where a general-domain NMT model is adapted to a specific domain by learning from the feedback of a human translator. This is a realistic interactive learning scenario where cost-free pre-training on a general domain data is possible, but each feedback generated by the human translator in the personalization step incurs a specific cost. In our experiment, we use human-generated reference translations to simulate both the cost of human feedback and to measure the performance gain achieved by model updates.

4.1 Experimental Setup

Seq2Seq Architecture. Both the Seq2Seq learner and the regulator are based on LSTMs (Hochreiter and Schmidhuber, 1997). The Seq2Seq has four bi-directional encoder and four decoder layers with 1024 units each, embedding layers of size 512. It uses Luong et al. (2015)'s input feeding and output layer, and global attention with a single feed forward layer (Bahdanau et al., 2015).

Regulator Architecture. The regulator consists of LSTMs on two levels: Inspired by Siamese Networks (Bromley et al., 1994), a bi-directional LSTM encoder of size 512 separately reads in both the current input sequence and the beam search hypothesis generated by the Seq2Seq. The last state of encoded source and hypothesis sequence and the previous output distribution are concatenated to form the input to a higher-level regulator LSTM of size 256. This LSTM updates its internal state and predicts a score for every feedback type for every input in the mini-batch. The feedback for each input is chosen by sampling from the distribution obtained by softmax normalization of these scores. The embeddings of the regulator are initialized by the Seq2Seq's source embeddings and further tuned during training. The model is implemented in the JoeyNMT[1] framework based on PyTorch.[2]

Data. We use three parallel corpora for German-to-English translation: a general-domain data set from the WMT2017 translation shared task for Seq2Seq pre-training, TED talks from the IWSLT2017 evaluation campaign for training the regulator with simulated feedback, and the Books

[1] https://github.com/joeynmt/joeynmt
[2] Code: https://github.com/juliakreutzer/joeynmt/tree/acl19

corpus from the OPUS collection (Tiedemann, 2012) for testing the regulator on another domain. Data pre-processing details and splits are given in §A.1. The joint vocabulary for Seq2Seq and the regulator consists of 32k BPE sub-words (Sennrich et al., 2016) trained on WMT.

Training. The Seq2Seq model is first trained on WMT with Adam (Kingma and Ba, 2015) on mini-batches of size 64, an initial learning rate 1×10^{-4} that is halved when the loss does not decrease for three validation rounds. Training ends when the validation score does not increase any further (scoring 29.08 BLEU on the WMT test). This model is then adapted to IWSLT with self-regulated training for one epoch, with online human feedback simulated from reference translations. The mini-batch size is reduced to 32 for self-regulated training to reduce the credit assignment problem for the regulator. The constant cost α (Eq. 1) is set to 1.[3] When multiple runs are reported, the same set of random seeds is used for all models to control the order of the input data. The best run is evaluated on the Books domain for testing the generalization of the regulation strategies.

Simulation of Cost and Performance. In our experiments, human feedback and its cost, and the performance gain achieved by model updates, is simulated by using human reference translations. Inspired by the keystroke mouse-action ratio (KSMR) (Barrachina et al., 2009), a common metric for measuring human effort in interactive machine translation, we define feedback cost as the sum of costs incurred by character edits and clicks, similar to Peris and Casacuberta (2018). The cost of a full correction (FULL) is the number of character edits between model output and reference, simulating the cost of a human typing.[4] Error markings (WEAK) are simulated by comparing the hypothesis to the reference and marking the longest common sub-strings as correct, as proposed by Petrushkov et al. (2018). As an extension to Petrushkov et al. (2018) we mark multiple common sub-strings as correct if all of them have the longest length. The cost is defined as the number of marked words, assuming an interface that allows markings by clicking on words. For self-training (SELF) and skipping training instances we naively assume zero cost, thus limiting the measurement of cost to the effort of the human teacher, and neglecting the effort on the learner's side. Table 1 illustrates the costs per feedback type on a randomly selected set of examples.

We measure the model improvement by evaluating the held-out set translation quality of the learned model at various time steps with corpus BLEU (cased `SacreBLEU` (Post, 2018)) and measure the accumulated costs. The best model is considered the one that delivers the highest quality at the lowest cost. This trade-off is important to bear in mind since it differs from the standard evaluation of machine translation models, where the overall best-scoring model, regardless of the supervision cost, is considered best. Finally, we evaluate the strategy learned by the regulator on an unseen domain, where the regulator decides which type of feedback the learner gets, but is not updated itself.

4.2 Results

We compare learning from one type of feedback in isolation against regulators with the following set of actions:

1. *Reg2*: FULL, WEAK
2. *Reg3*: FULL, WEAK, SELF
3. *Reg4*: FULL, WEAK, SELF, NONE

Cost vs. Quality. Figure 3 compares the improvement in corpus BLEU (Papineni et al., 2002) (corresponding to results in Translation Error Rate (TER, computed by `pyTER`) (Snover et al., 2006)) of regulation variants and full feedback over cumulative costs of up to 80k character edits. Using only full feedback (blue) as in standard supervised learning or learning from post-edits, the overall highest improvement can be reached (visible only after the cutoff of 80k edits; see Appendix A.2 for the comparison over a wider window of time). However, it comes at a very high cost (417k characters in total to reach +0.6 BLEU). The regulated variants offer a much cheaper improvement, at least until a cumulative cost between 80k (*Reg4*) and 120k (*Reg2*), depending on the feedback options available. The regulators do not reach the quality of the full model since their choice of feedback is oriented towards costs and immediate improvements. By finding a trade-off between feedback types for immediate improvements, the regulators sacrifice long-term improvement. Comparing regulators, *Reg2* (orange) reaches the overall

[3] Values $\neq 1$ distort the rewards for self-training too much.
[4] As computed by the Python library `difflib`.

SELF	0	x	Sie greift in ihre Geldbörse und gibt ihm einen Zwanziger .
		$\hat{y}$	It attacks their wallets and gives him a twist .
		y^*	She reaches into her purse and hands him a 20 .
WEAK	9	x	Und als ihr Vater sie sah und sah , wer sie geworden ist , in ihrem vollen Mädchen-Sein , schlang er seine Arme um sie und brach in Tränen aus .
		$\hat{y}$	And when her father saw them and saw who became them , in their full girl 's , he swallowed his arms around them and broke out in tears .
		y^*	When her father saw her and saw who she had become , in her full girl self , he threw his arms around her and broke down crying .
FULL	59	x	Und durch diese zwei Eigenschaften war es mir möglich , die Bilder zu erschaffen , die Sie jetzt sehen .
		$\hat{y}$	And ~~through~~ these two ~~features ,~~ I was able to create the images you now ~~see~~ .
		y^*	And it was with those two properties that I was able to create the images that you 're seeing right now .

Table 1: Examples from the IWSLT17 training set, cost (2nd column) and feedback decisions made by *Reg3*. For weak feedback, marked parts are underlined, for full feedback, the corrections are marked by underlining the parts of the reference that got inserted and the parts of the hypothesis that got deleted.

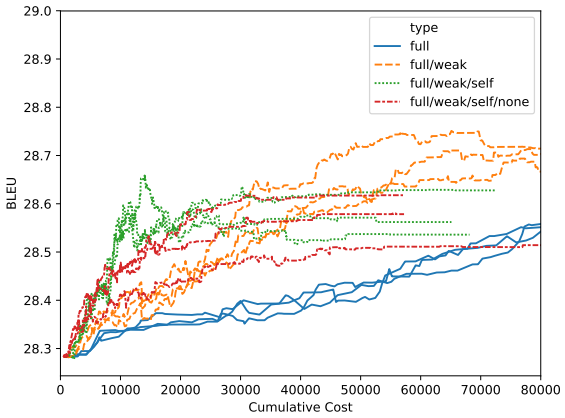


Figure 3: BLEU of regulation variants over cumulative costs. BLEU is computed on the tokenized IWSLT validation set with greedy decoding.

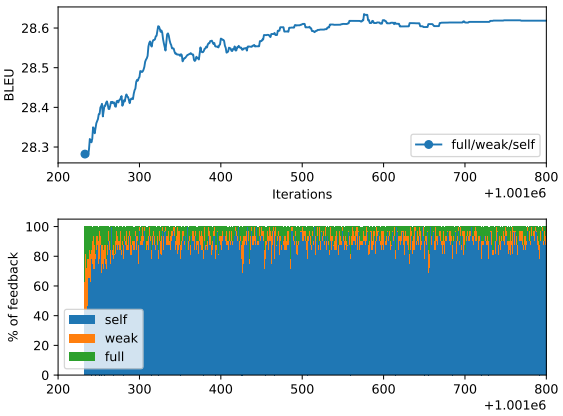


Figure 4: *Reg3* actions as chosen over time, depicted for each iteration. Counting of iterations starts at the previous iteration count of the baseline model.

highest improvement over the baseline model, but until the cumulative cost of around 35k character edits, *Reg3* (green) offers faster improvement at a lower cost since it has an additional, cheaper feedback option. Adding the option to skip examples (*Reg4*, red) does not give a benefit. Appendix A.3 lists detailed results for offline evaluation on the trained Seq2Seq models on the IWSLT test set: Self-regulating models achieve improvements of 0.4-0.5 BLEU with costs reduced up to a factor of 23 in comparison to the full feedback model. The reduction in cost is enabled by the use of cheaper feedback, here markings and self-supervision, which in isolation are very successful as well. Self-supervision works surprisingly well and can be recommended for cheap but effective unsupervised domain adaptation for sequence-to-sequence learning.

Self-Regulation Strategies. Figure 4 shows which actions *Reg3* chooses over time when trained on IWSLT. Most often it chooses to do self-training on the current input. The choice of feedback within one batch varies only slightly during training, with the exception of an initial exploration phase within the first 100 iterations. In general, we observe that all regulators are highly sensitive to balancing cost and performance, and mostly prefer the cheapest option (e.g., *Reg4* by choosing mostly NONE) since they are penalized heavily for choosing (or exploring) expensive options (see Eq. 1).

A further research question is whether and how the self-regulation module takes the input or output context into account. We therefore compare its decisions to a context-free ϵ-greedy strategy. The ϵ-greedy algorithm is a successful algorithm for multi-armed bandits (Watkins, 1989). In our case, the arms are the four feedback types. They are chosen based on their reward statistics, here the average empirical reward per feedback type $Q_i(s) = \frac{1}{N_i(s)} \sum_{0,\ldots,i} r(s_i)$. With probability $1-\epsilon$, the algorithm selects the feedback type with the highest empirical reward (exploitation), otherwise picks one of the remaining arms at random (exploration). In contrast to the neural regulator model, ϵ-greedy decides solely on the basis

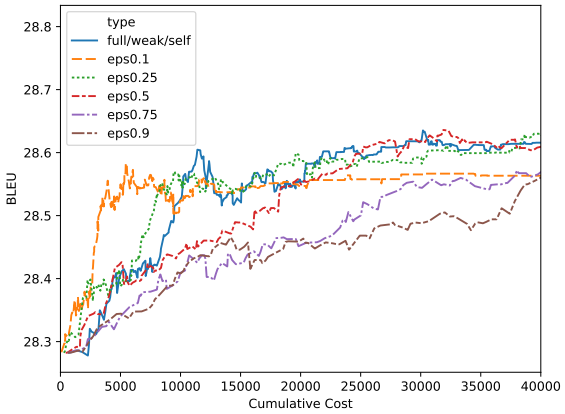


Figure 5: BLEU and cumulative costs on IWSLT for *Reg3* and ϵ-greedy with $\epsilon \in [0.1, 0.25, 0.5, 0.75, 0.9]$.

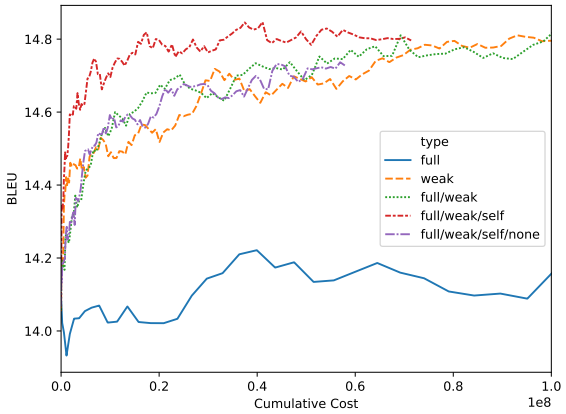


Figure 6: Domain transfer of regulators trained on IWSLT to the Books domain in comparison to full and weak feedback only.

of the reward statistics and has no internal contextual state representation. The comparison of *Reg3* with ϵ-greedy for a range of values for ϵ in Figure 5 shows that learned regulator behaves indeed very similar to an ϵ-greedy strategy with $\epsilon = 0.25$. ϵ-greedy variants with higher amounts of exploration show a slower increase in BLEU, while those with more exploitation show an initial steep increase that flattens out, leading to overall lower BLEU scores. The regulator has hence found the best trade-off, which is an advantage over the ϵ-greedy algorithm where the ϵ hyperparameter requires dedicated tuning. Considering the ϵ-greedy-like strategy of the regulator and the strong role of the cost factor shown in Figure 4, the regulator module does not appear to choose individual actions based e.g., on the difficulty of inputs, but rather composes mini-batches with a feedback ratio according to the feedback type's statistics. This confirms the observations of Peris and Casacuberta (2018), who find that the subset of instances selected for labeling is secondary—it is rather the mixing ratio of feedback types that matters. This finding is also consistent with the mini-batch update regime that forces the regulator to take a higher-level perspective and optimize the expected improvement at the granularity of (mini-batch) updates rather than at the input level.

Domain Transfer. After training on IWSLT, we evaluate the regulators on the Books domain: Can they choose the best actions for an efficient learning progress without receiving feedback on the new domain? We evaluate the best run of each regulator type (i.e., ϕ trained on IWSLT), with the Seq2Seq model reset to the WMT baseline. The regulator is not further adapted to the Books domain, but decides on the feedback types for training the Seq2Seq model for a single epoch on the Books data. Figure 6 visualizes the regulated training process of the Seq2Seq model. As before, *Reg3* performs best, outperforming weak, full and self-supervision (reaching 14.75 BLEU, not depicted since zero cost). Learning from full feedback improves much later in training and reaches 14.53 BLEU.[5] One explanation is that the reference translations in the Books corpus are less literal than the ones for IWSLT, such that a weak feedback signal allows the learner to learn more efficiently than from full corrections. Appendix A.4 reports the results for offline evaluation on the trained Seq2Seq models on the Books test set.

Comparison to Active Learning. A classic active learning strategy is to sample a subset of the input data for full labeling based on the uncertainty of the model predictions (Settles and Craven, 2008). The size of this subset, i.e. the amount of human labeling effort, has to be known and determined before learning. Figure 7 compares the self-regulators on the Books domain with models that learn from a fixed ratio of fully-labeled instances in every batch. These are chosen according to the model's uncertainty, here measured by the average token entropy of the model's best-scoring beam search hypothesis. The regulated models with a mix of feedback types clearly outperform the active learning strategies,

[5] With multiple epochs it would improve further, but we avoid showing the human the same inputs multiple times.

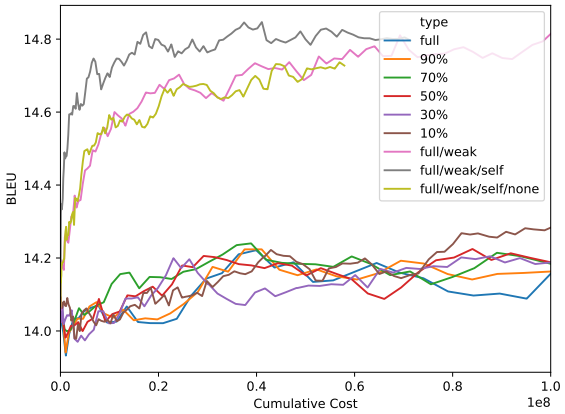


Figure 7: Learned self-regulation strategies in comparison to uncertainty-based active learning with a fixed percentage of full feedback on the Books domain.

both in terms of cost-efficient learning (Figure 7) as well as in overall quality (See Figure 9 in Appendix A.5). We conclude that mixing feedback types, especially in the case where full feedback is less reliable, offers large improvements over standard stream-based active learning strategies.

4.3 Prospects for Field Studies

Our experiments were designed as a pilot study to test the possibilities of self-regulated learning in simulation. In order to advance to field studies where human users interact with Seq2Seq models, several design choices have to be adapted with caution. Firstly, we simulate both feedback cost and quality improvement by measuring distances to static reference outputs. The experimental design in a field study has to account for a variation of feedback strength, feedback cost, and performance assessments, across time, across sentences, and across human users (Settles et al., 2008). One desideratum for field studies is thus to analyze this variation by analyzing the experimental results in a mixed effects model that accounts for variability across sentences, users, and annotation sessions (Baayen et al., 2008; Karimova et al., 2018). Secondly, our simulation of costs considers only the effort of the human teacher, not the machine learner. The strong preference for the cheapest feedback option might be a result of overestimating the cost of human post-editing and underestimating the cost of self-training. Thus, a model for field studies where data is limited might greatly benefit from learned estimates of feedback cost and quality improvement (Kreutzer et al., 2018).

5 Conclusion

We proposed a cost-aware algorithm for interactive sequence-to-sequence learning, with a self-regulation module at its core that learns which type of feedback to query from a human teacher. The empirical study on interactive NMT with simulated human feedback showed that this self-regulated model finds more cost-efficient solutions than models learning from a single feedback type and uncertainty-based active learning models, also under domain shift. While this setup abstracts away from certain confounding variables to be expected in real-life interactive machine learning, it should be seen as a pilot experiment that allows focussing on our central research questions under an exact and noise-free computation of feedback cost and performance gain. The proposed framework can naturally be expanded to integrate more feedback modes suitable for the interaction with humans, e.g., pairwise comparisons or output rankings. Future research directions will involve the development of reinforcement learning model with multi-dimensional rewards, and modeling explicit credit assignment for improving the capabilities of the regulator to make context-sensitive decisions in mini-batch learning.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback. The research reported in this paper was supported in part by the German research foundation (DFG) under grant RI-2221/4-1.

References

R Harald Baayen, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations (ICLR)*, San Diego, California, USA.

Sergio Barrachina, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and Juan-Miguel Vilar. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1).

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, Denver, CO, USA.

Carlos Celemin, Javier Ruiz-del Solar, and Jens Kober. 2019. A fast hybrid reinforcement learning framework with human corrective feedback. *Autonomous Robots*, 43(5):1173–1186.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Ching-An Cheng, Xinyan Yan, Nolan Wagener, and Byron Boots. 2018. Fast policy learning through imitation and reinforcement. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, Monterey, CA, USA.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium.

Miguel Domingo, Álvaro Peris, and Francisco Casacuberta. 2017. Segment-based interactive-predictive machine translation. *Machine Translation*, 31(4):163–185.

Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to active learn: A deep reinforcement learning approach. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning (ICML)*, Vancouver, Canada.

John Hattie and Gregory M. Donoghue. 2016. Learning strategies: a synthesis and conceptual model. *NPJ Science of Learning*, 1:16013–16013.

John Hattie and Helen Timperley. 2007. The power of feedback. *American Educational Research Association*, 77(1):81–112.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

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

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. 2017. Reinforcement learning with unsupervised auxiliary tasks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France.

Sariya Karimova, Patrick Simianer, and Stefan Riezler. 2018. A user-study on online adaptation of neural machine translation to human post-edits. *Machine Translation*, 32(4):309–324.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

David Krueger, Jan Leike, Owain Evans, and John Salvatier. 2016. Active reinforcement learning: Observing rewards at a cost. In *Proceeding of the 30th Conference on Neural Information Processing Systems (NeurIPS)*, Barcelona, Spain.

Gaurav Kumar, George Foster, Colin Cherry, and Maxim Krikun. 2019. Reinforcement learning based curriculum optimization for neural machine translation. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Minneapolis, MN, USA.

Tsz Kin Lam, Julia Kreutzer, and Stefan Riezler. 2018. A reinforcement learning approach to interactive-predictive neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT)*, Alicante, Spain.

Ming Liu, Wray Buntine, and Gholamreza Haffari. 2018. Learning to actively learn neural machine translation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL)*, Brussels, Belgium.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

James MacGlashan, Mark K. Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. 2017. Interactive learning from policy-dependent human feedback. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Sydney, Australia.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2015. Never-ending learning. In *Proceedings of the 29th Conference on Artificial Intelligence (AAAI)*, Austin, TX, USA.

Joel T. Nigg. 2017. Annual research review: On the relations among self-regulation, self-control, executive functioning, effortful control, cognitive control, impulsivity, risk-taking, and inhibition for developmental psychopathology. *Journal of Child Psychology and Psychiatry*, 58(4):361–383.

Ernesto Panadero. 2017. A review of self-regulated learning: Six models and four directions of research. *Frontiers in Psychology*, 8(422):1–28.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA.

Álvaro Peris and Francisco Casacuberta. 2018. Active learning for interactive neural machine translation of data streams. In *Proceedings of the 22nd Conference on Computational Natural Language Learning (CONLL)*, Brussels, Belgium.

Pavel Petrushkov, Shahram Khadivi, and Evgeny Matusov. 2018. Learning from chunk-based feedback in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT)*, Brussels, Belgium.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representation (ICLR)*, San Juan, Puerto Rico.

Jürgen Schmidhuber, Jieyu Zhao, and Marco Wiering. 1996. Simple principles of metalaerning. Technical Report 69 96, IDSIA, Lugano, Switzerland.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Honolulu, Hawaii.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of the NeurIPS Workshop on Cost-Sensitive Learning*, Vancouver, Canada.

Matthew J. A. Smith, Herke Van Hoof, and Joelle Pineau. 2018. An inference-based policy gradient method for learning options. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas (AMTA)*, volume 200.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NeurIPS)*, Montreal, Canada.

Sebastian Thrun and Lorien Pratt, editors. 1998. *Learning to Learn*. Kluwer, Dortrecht, MA, USA.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Marco Turchi, Matteo Negri, M Amin Farajian, and Marcello Federico. 2017. Continuous learning from human post-edits for neural machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):233–244.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA.

Christopher Watkins. 1989. Learning from delayed rewards. *PhD thesis, Cambridge University*.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

Lijun Wun, Fei Tian, Yingce Xia, Yang Fan, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. Learning to teach with dynamic loss functions. In *Proceeding of the 32nd Conference on Neural Information Processing System (NeuRIPS)*, Montreal, Canada.

Barry J. Zimmerman and Dale H. Schunk, editors. 1989. *Self-Regulated Learning and Academic Achievement*. Springer, New York, NY, USA.

A Appendices

A.1 Data

de↔en	WMT	IWSLT	Books
Train	5,889,699	206,112	46,770
Dev	2,169	2,385	2,000
Test	3,004	1,138	2,000

Table 2: Number of sentences for parallel corpora used for pre-training (WMT), regulator training (IWSLT) and domain transfer evalution (Books).

The WMT data is obtained from the WMT 2017 shared task website[6] and pre-processed as described in Hieber et al. (2017). The pre-processing pipeline is used for IWSLT and Books data as well. IWSLT2017 is obtained from the evaluation campaign website.[7] For validation on WMT, we use the `newstest2015` data, for IWSLT `tst2014+tst2015`, for testing on WMT `newstest2017` and `tst2017` for IWSLT. Since there is no standard split for the Books corpus, we randomly select 2k sentences for validation and testing each. Table 2 gives an overview of the size of the three resources.

A.2 Online Evaluation on IWSLT

Figure 8 displays the development of BLEU over costs and time.

A.3 Offline Evaluation on IWSLT

Table 3 reports the offline held-out set evaluations for the early stopping points selected on the dev set for all feedback modes. All models notably improve over the baseline, only using full feedback leads to the overall best model on IWSLT (+0.6 BLEU / -0.6 TER), but costs a massive amounts of edits (417k characters). Self-regulating models still achieve improvements of 0.4–0.5 BLEU/TER with costs reduced up to a factor of 23. The reduction in cost is enabled by the use of cheaper feedback, here markings and self-supervision, which in isolation are successful as well. Self-supervision works surprisingly well, which makes it attractive for cheap but effective unsupervised domain adaptation. It has to be noted that both weak and self-supervision worked only well when targets were pre-computed with the baseline model and held fixed during training. We suspect that the strong reward signal ($f_t = 1$) for non-reference outputs leads otherwise to undesired local overfitting effects that a learner with online-generated targets cannot recover from.

[6] http://www.statmt.org/wmt17/translation-task.html

[7] https://sites.google.com/site/iwsltevaluation2017/

Model	IWSLT dev		IWSLT test	
	BLEU↑	Cost↓	BLEU↑	TER↓
Baseline	28.28	-	24.84	62.42
Full	$28.93_{\pm 0.02}$	417k	$25.60_{\pm 0.02}$	$61.86_{\pm 0.03}$
Weak	$28.65_{\pm 0.01}$	32k	$25.10_{\pm 0.09}$	$62.12_{\pm 0.12}$
Self	$28.58_{\pm 0.02}$	-	$25.33_{\pm 0.06}$	$61.96_{\pm 0.05}$
Reg4	$28.57_{\pm 0.04}$	68k	$25.23_{\pm 0.05}$	$62.02_{\pm 0.12}$
Reg3	$28.61_{\pm 0.03}$	18k	$25.23_{\pm 0.09}$	$62.07_{\pm 0.06}$
Reg2	$28.66_{\pm 0.06}$	88k	$25.27_{\pm 0.09}$	$61.91_{\pm 0.06}$

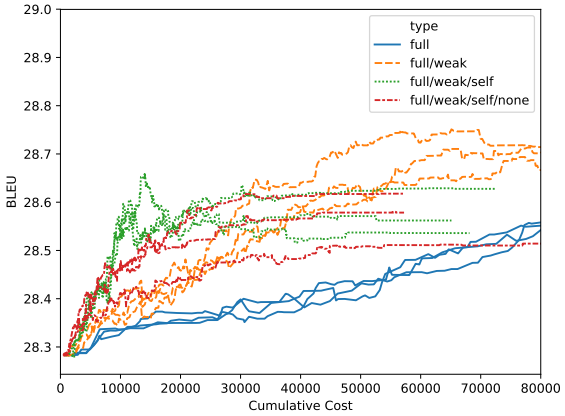
Table 3: Evaluation of models at early stopping points. Results for three random seeds on IWSLT are averaged, reporting the standard deviation in the subscript. The translation of the dev set is obtained by greedy decoding (as during validation) and of the test set with beam search of width five. The costs are measured in character edits and clicks, as described in Section 4.

A.4 Domain Transfer

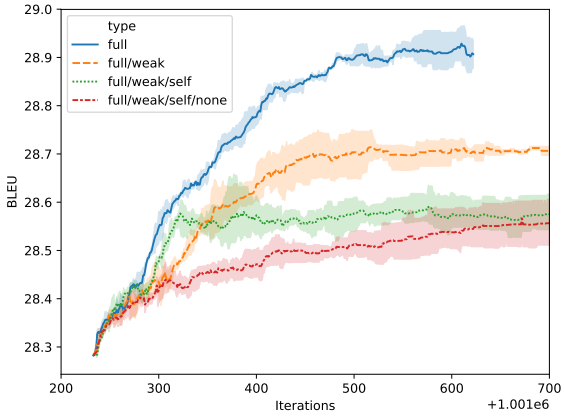
Model	Books test		
	BLEU↑	TER↓	Cost↓
Baseline	14.19	79.81	-
Full	14.87	79.12	1B
Weak	14.74	78.14	93M
Self	14.73	78.86	-
Reg4	14.80	79.02	57M
Reg3	14.80	78.70	41M
Reg2	15.00	78.21	142M

Table 4: Evaluation of models at early stopping points on the Books test set (beam search with width five).

Table 4 reports results for test set evaluation on the Books domain of the best model from the IWSLT domain each. The baseline was trained on WMT parallel data without any regulation. The regulator was trained on IWSLT and evaluated on Books, the Seq2Seq model is further trained for one epoch on Books. The costs are measured in character edits and clicks. The best result in terms of BLEU and TER is achieved by the *Reg2* model, even outperforming the model with full feedback. As observed for the IWSLT domain (cf. Section 4.2), self-training is very effective, but is outperformed by the *Reg2* model and roughly on par



(a) BLEU over cumulative costs.



(b) BLEU over time.

Figure 8: Regulation variants evaluated in terms of BLEU over time (a) and cumulative costs (b). Iteration counts start from the iteration count of the baseline model. One iteration on IWSLT equals training on one mini-batch of 32 instances. The BLEU score is computed on the tokenized validation set with greedy decoding. In (b) the lines correspond to the means over three runs, the shaded area depicts the estimated 95% confidence interval.

with the *Reg3* model.

A.5 Active Learning on Books

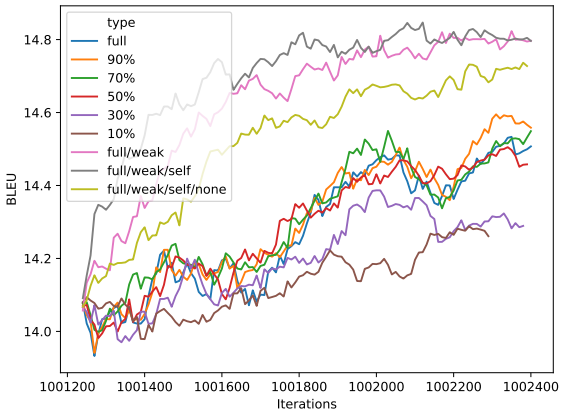


Figure 9: Development of validation BLEU over time for learned regulation strategies in comparison to active learning with a fixed percentage γ of full feedback. Counting of iterations starts at the previous iteration count of the baseline model.

Figure 9 shows the development of BLEU over time for the regulators and active learning strategies with a fixed ratio of full feedback per batch ($\gamma \in [10, 30, 50, 70, 90]$). The decision whether to label an instance in a batch is made based on the average token entropy of the model's current hypothesis. Using only 50% of the fully-supervised labels achieves the same quality as 100% using this uncertainty-based active learning sampling strategy. However, the regulated models reach a higher quality not only at a lower cost (see Figure 7), but also reach an overall higher quality.

A.6 Regulation Strategies on IWSLT

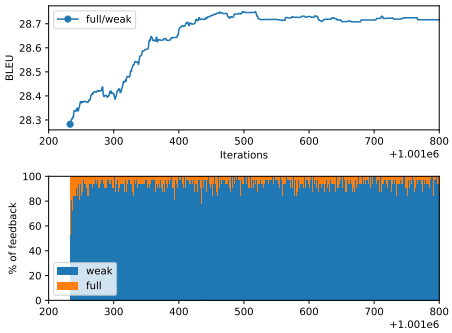


Figure 10: Feedback chosen by *Reg2* on IWSLT.

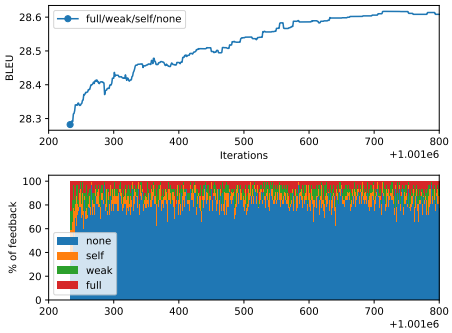


Figure 11: Feedback chosen by *Reg4* on IWSLT.

Figures 10 and 11 show the ratio of feedback types for self-regulation during training with *Reg2* and *Reg4* respectively.