A Reinforcement Learning Approach to Interactive-Predictive Neural Machine Translation

Tsz Kin Lam[†,*] and **Julia Kreutzer**[*] and **Stefan Riezler**[†,*]
[*]Computational Linguistics & [†]IWR, Heidelberg University, Germany
{lam,kreutzer,riezler}@cl.uni-heidelberg.de

Abstract

We present an approach to interactive-predictive neural machine translation that attempts to reduce human effort from three directions: Firstly, instead of requiring humans to select, correct, or delete segments, we employ the idea of learning from human reinforcements in form of judgments on the quality of partial translations. Secondly, human effort is further reduced by using the entropy of word predictions as uncertainty criterion to trigger feedback requests. Lastly, online updates of the model parameters after every interaction allow the model to adapt quickly. We show in simulation experiments that reward signals on partial translations significantly improve character F-score and BLEU compared to feedback on full translations only, while human effort can be reduced to an average number of 5 feedback requests for every input.

1 Introduction

Interactive-predictive machine translation aims at obtaining high-quality machine translation by involving humans in a loop of user validations of partial translations suggested by the machine translation system. This interaction protocol can easily be fit to neural machine translation (NMT) (Bahdanau et al., 2015) by conditioning the model's word predictions on the user-validated prefix (Knowles and Koehn, 2016; Wuebker et al., 2016). User studies conducted by Green et al. (2014) for phrase-based machine translation have shown that the interactive-predictive interaction protocol leads to significant reductions in post-editing effort. Other user studies on interactive machine translation based on post-editing have shown that human effort can also be reduced by improving the online adaptation capabilities of the learning system, both for statistical phrase-based (Bentivogli et al., 2016) or NMT systems (Karimova et al., 2017).

The goal of our work is to further reduce human effort in interactive-predictive NMT by combining the advantages of the interactive-predictive protocol with the advantages of learning from weak feedback. For the latter we rely on techniques from reinforcement learning (Sutton and Barto, 2017), a.k.a. bandit structured prediction (Sokolov et al., 2016; Kreutzer et al., 2017; Nguyen et al., 2017) in the context of sequence-to-sequence learning. Our approach attacks the problem of reducing human effort from three innovative directions.

- Firstly, instead of requiring humans to correct or delete segments proposed by the machine translation system, we employ the reinforcement learning idea of humans providing reward signals in form of judgments on the quality of the machine translation. Human effort is reduced since each partial translation receives a human reward signal at most once, rendering it a bandit-type feedback signal, and each reward signal itself is easier to obtain than a correction of a translation.
- In order to reduce the amount of feedback signals even further, we integrate an uncertainty criterion for word predictions to trigger requests for human feedback. Using the comparison of the current average entropy to

the entropy of word predictions in the history as a measure for uncertainty, we reduce the amount of feedbacks requested from humans to an average number of 5 requests per input.

- In contrast to previous approaches to interactive-predictive translation, the parameters of our translation system are updated online after receiving feedback for partial translations. The update is done according to an actor-critic reinforcement learning protocol where each update pushes up the score function of the partial translation sampled by the model (called actor) proportional to a learned reward function (called critic). Furthermore, since the entropy criterion is based on the actor, it is also automatically updated. Frequent updates improve the adaptability of our system, resulting in a further reduction of human effort.

The rest of this paper is structured as follows. In Section 2, we will situate our approach in the context of interactive machine translation and analyze our contribution related to reinforcement learning for sequence prediction problems. Details of our algorithm are given in Section 3. We evaluate our approach in a simulation study where bandit feedback is computed by evaluating partial translations against references under a character F-score metric (Popović, 2015) without revealing the reference translation to the learning system (Section 4). We show that segment-wise reward signals improve translation quality over reinforcement learning with sparse sentence-wise rewards, measured by character F-score and corpus-based BLEU against references. Furthermore, we show that human effort, measured by the number of feedback requests, can be reduced to an average number of 5 requests per input. These implications of our new paradigm are discussed in Section 5.

2 Related Work

The interactive-predictive translation paradigm reaches back to early approaches for IBM-type (Foster et al., 1997; **?**) and phrase-based machine translation (Barrachina et al., 2008; Green et al., 2014). Knowles and Koehn (2016) and Wuebker et al. (2016) presented *neural interactive translation prediction* — a translation scenario where translators interact with an NMT system by accepting or correcting subsequent target tokens suggested by the NMT system in an auto-complete style. NMT is naturally suited for this incremental production of outputs, since it models the probability of target tokens given a history of target tokens sequentially from left to right. In standard supervised training with teacher forcing, this history comes from the ground truth, while in interactive-predictive translation it is provided by the prefix accepted or entered by the user. Both approaches use references to simulate an interaction with a translator and compare their approach to phrase-based prefix-search. They find that NMT is more accurate in word and letter prediction and recovers better from failures. Similar to their work, we will experiment in a simulated environment with references mimicking the translator. However, we do not use the reference directly for teacher forcing, but only to derive weak feedback from it. Furthermore, our approach employs techniques to reduce the number of interactions, and to update the model more frequently than after each sentence.

Our work is also closely related to approaches for *interactive pre-post-editing* (Marie and Max, 2015; Domingo et al., 2018). The core idea is to ask the translator to mark good segments and use these for a more informed re-decoding. Both studies could show a reduction in human effort for post-editing in simulation experiments. We share the goal of using human feedback more effectively by targeting it towards essential translation segments, however, our approach does adhere to the left-to-right navigation through translation hypotheses. In difference to these approaches, we try to reduce human effort even further by minimizing the number of feedback requests and by frequent model updates.

Reinforcing/penalizing a targeted set of actions can also be found in recent approaches to *reinforcement learning from human feedback*. For example, Judah et al. (2010) presented a scenario where users interactively label freely chosen good and bad parts of a policy's trajectory. The policy is directly trained with this reinforcement signal to play a real-time strategy game. Simulations of NMT systems interacting with human feedback have been presented firstly by Kreutzer et al. (2017), Nguyen (2017), or Bahdanau et al. (2017) who apply different policy gradient algorithms, William's REINFORCE (Williams, 1992) or actor-critic methods (Konda and Tsitsiklis, 2000; Sutton et al., 2000; Mnih et al., 2016), respectively. While

Bahdanau et al.'s (2017) approach operates in a fully supervised learning scenario, where rewards are simulated in comparison to references with smoothed and length-rescaled BLEU, Kreutzer et al. (2017) and Nguyen et al. (2017) limit the setup to sentence-level bandit feedback, i.e. only one feedback is obtained for one completed translation per input. In this paper, we use actor-critic update strategies, but we receive simulated bandit feedback on the sub-sentence level.

We adopt techniques from *active learning* to reduce the number of feedbacks requested from a user. González-Rubio et al. (2011; 2012) apply active learning for interactive machine translation, where a user interactively finishes the translation of an SMT system. The active learning component decides which sentences to sample for translation (i.e. receive full supervision for) and the SMT system is updated online (Ortiz-Martínez et al., 2010). In our algorithm the active learning component decides which prefixes to be rated (i.e. receive weak feedback for) based on their average entropy. Entropy is a popular measure for uncertainty in active learning: the rationale is to feed the learning algorithm with labeled instances where it is least confident about its own predictions. This *uncertainty sampling* algorithm (Lewis and Gale, 1994) is a popular choice for active learning for NLP tasks with expensive gold labeling, such as text classification (Lewis and Gale, 1994), word-sense disambiguation (Chen et al., 2006) and statistical parsing (Tang et al., 2002). Our method falls into the category of stream-based online active learning (as opposed to pool-based active learning, selecting instances from a large pool of unlabeled data), since the algorithm decides on the fly (online) which translation prefixes of the stream of source tokens to request feedback for. Instead of receiving gold annotations, as in the studies mentioned above, our algorithm receives weaker, bandit feedback — but the motivation of minimizing human labeling effort is the same.

3 Reinforcement Learning for Interactive-Predictive Translation

In the following, we will introduce the key ideas of our approach, formalize them, and present an algorithm for reinforcement learning for interactive-predictive NMT.

3.1 Actor-Critic Reinforcement Learning for NMT

The objective of reinforcement learning methods is to maximize the expected reward obtainable from interactions of an agent (here: a machine translation system) with an environment (here: a human translator). In our case, the agent/system performs actions by predicting target words y_t according to a stochastic policy p_θ parameterized by an RNN encoder-decoder NMT system (Bahdanau et al., 2015) where

$$p_\theta(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T_y} p_\theta(y_t|\mathbf{x}, \mathbf{y}_{<t}). \tag{1}$$

The environment/human can be formalized as a Markov Decision Process where a state at time t is a tuple $s_t = \langle \mathbf{x}, \mathbf{y}_{<t} \rangle$ consisting of the conditioning context of the input $\mathbf{x}$ and the current produced history of target tokens $\mathbf{y}_{<t}$. Note that since states s_{t+1} include the current chosen action y_t and can contain long histories $\mathbf{y}_{<t}$, the state distribution is sparse and deterministic. The reward distribution of the environment/critic is estimated by function approximation in actor-critic methods. The reward estimator (called critic) is trained on actual rewards and updated after every interaction, and then used to update the parameters of the policy (called actor) in a direction of function improvement. We use the advantage actor critic framework of Mnih et al. (2016) which estimates the advantage $A_\phi(y_t|s_t)$ in reward of choosing action y_t in a given state s_t over the mean reward value for that state. This framework has been applied to reinforcement learning for NMT by Nguyen et al. (2017). The main objective of the actor is then to maximize the expected advantage

$$L_\theta = \mathbb{E}_{p(\mathbf{x})p_\theta(\mathbf{y}|\mathbf{x})} \left[\sum_{t=1}^{T_y} A_\phi(y_t|s_t) \right]. \tag{2}$$

The stochastic gradient of this objective for a sampled target word $\hat{y}_t$ for an input $\mathbf{x}$ can be calculated following the policy gradient theorem (Sutton et al., 2000; Konda and Tsitsiklis, 2000) as

$$\nabla L_\theta(\hat{y}_t) = \sum_{t=1}^{T_y} \left[\nabla \log p_\theta(\hat{y}_t|s_t) A_\phi(\hat{y}_t|s_t) \right]. \tag{3}$$

In standard actor-critic algorithms, the parameters of actor and the critic are updated online at each time step. The actor parameters θ are updated by sampling $\hat{y}_t$ from p_θ and performing a step in the opposite direction of the stochastic gradient of $L_\theta(\hat{y}_t)$; the critic parameters ϕ are updated by minimizing $L_\phi(\hat{y}_t)$, defined as the mean squared error of the reward estimator for sampled target word $\hat{y}_t$ with respect to actual rewards (for more details see Nguyen et al. (2017)). In our experiments, we simulate user rewards by character F-score (chrF) values of partial translations.

3.2 Triggering Human Feedback Requests by Actor Entropy

Besides the idea of replacing human post-edits by human rewards, another key feature of our approach is to minimize the number of requests for human feedback. This is achieved by computing the uncertainty of the policy distribution as the average word-level entropy $\bar{H}$ of an n-word partial translation, defined as

$$\bar{H}(\hat{y}_{1:n}) = \frac{1}{n} \sum_{t=1}^{n} \left[- \sum_{v \in \mathcal{V}} p_\theta(v|s_t) \log p_\theta(v|s_t) \right], \tag{4}$$

where $\hat{y}_{1:n} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$ is a sequence of n predicted tokens starting at the sentence beginning, $\mathcal{V}$ is the output vocabulary, and $p_\theta(v|s_t)$ is the probability of predicting a word in $\mathcal{V}$ at state s_t of the RNN decoder.

A request for human feedback is triggered when $\bar{H}(\hat{y}_{1:n})$ is higher than a running average γ by a factor of ϵ or when <eos> is generated. Upon receiving a reward from the user, both actor and critic are updated. Hence, our algorithm takes the middle ground between updating at each time step t and performing an update only after a reward signal for the completed translation is received. In our simulation experiments, this process is repeated until the <eos> token is generated, or when a pre-defined maximum length, here $T_{\max} = 50$, is reached.

3.3 Simulating Human Rewards on Translation Quality

Previous work on reinforcement learning in machine translation has simulated human bandit feedback by evaluating full-sentence translations against references using per-sentence approximations of BLEU (Sokolov et al., 2016; Kreutzer et al., 2017; Nguyen et al., 2017). We found that when working with partial translations, user feedback on translation quality can successfully be simulated by computing the chrF-score (Popović, 2015) of the translation with respect to the reference translation truncated to the same length. If the length of the translation exceeds the length of the reference, no truncation is used. We denote rewards as a function $\mathrm{R}(\hat{y}_{1:t})$ of only the partial translation $\hat{y}_{1:t}$, in order to highlight the fact that rewards are in principle independent of reference translations.

3.4 Sampling versus Forced Decoding via Prefix Buffer Ξ

The standard approach to estimate the expected reward in policy gradient techniques is to employ Monte-Carlo methods, in specific, multinomial sampling of actions. This guarantees an unbiased estimator and allows sufficient exploration of the action space in learning. In contrast, interactive-predictive machine translation usually avoids exploration in favor of exploitation by decoding the best partial translation under the current model after every interaction. Since in our framework, learning and decoding are interleaved, we have to find the best compromise between exploration and exploitation.

The general modus operandi of our framework is simultaneous exploration and exploitation by multinomial sampling actions from the current policy. However, in cases where a partial translation receives a high user reward, we store it in a so-called prefix buffer Ξ, and perform forced decoding by feeding the prefix to the decoder for the remaining translation process.

3.5 Algorithm for Bandit Interactive-Predictive NMT

Algorithm 1 gives pseudo-code for **B**andit-**I**nteractive-**P**redictive **N**eural **M**achine **T**ranslation (BIP-NMT). The algorithm receives an input source sequence $\mathbf{x_i}$ (line 4), and incrementally predicts a sequence of output target tokens up to length $T_{\max}$ (line 6). At each step t, a partial translation $\hat{y}_{1:t}$ is sampled from the policy distribution $p_\theta(\cdot|\mathbf{x_i}, \mathbf{y}_{<t}, \Xi)$ that implements an RNN encoder-decoder with an additional prefix buffer Ξ for forced decoding (line 7). User feedback is requested in case the average entropy $\bar{H}(\hat{y}_{1:t})$ of the policy is larger than or equal to a running average by a factor of ϵ or when <eos> is generated (line 9). If the reward $\mathrm{R}(\hat{y}_{1:t})$ is larger than or equal to a threshold μ, the prefix is stored in a buffer for forced decoding (lines 11-12). Next,

Algorithm 1: Algorithm BIP-NMT

```
1:  Input: $\theta_0$, $\phi_0$, $\alpha_A$, $\alpha_C$
2:  Output: Estimates $\theta^*$, $\phi^*$
3:  for i = 1, ... N do
4:      Receive $\mathbf{x_i}$
5:      Initialize $\gamma \leftarrow 0$, $\Xi \leftarrow \emptyset$
6:      for t = 1 ... $T_{\max}$ do
7:          Sample $\hat{y}_{1:t} \sim p_{\theta_{t-1}}(\cdot|\mathbf{x_i}, \mathbf{y}_{<t}, \Xi)$
8:          Compute $\bar{H}(\hat{y}_{1:t})$ using Eq. (4)
9:          if $\bar{H}(\hat{y}_{1:t}) - \gamma_{t-1} \geq \epsilon \times \gamma_{t-1}$ or <eos>
            in $\hat{y}_{1:t}$ then
10:             Receive feedback R($\hat{y}_{1:t}$)
11:             if R($\hat{y}_{1:t}$) $\geq \mu$ then
12:                 $\Xi \leftarrow \hat{y}_{1:t}$
13:             end if
14:             Update $\theta_t \leftarrow \theta_{t-1} - \alpha_A \nabla L_{\theta_{t-1}}(\hat{y}_t)$
                (Eq. (3))
15:             Update $\phi_t \leftarrow \phi_{t-1} - \alpha_C \nabla L_{\phi_{t-1}}(\hat{y}_t)$
                (see Eq. (7) in Nguyen et al. (2017))
16:         end if
17:         Update $\gamma_t = \gamma_{t-1} + \frac{1}{t}\left(\bar{H}(\hat{y}_{1:t}) - \gamma_{t-1}\right)$
18:         break if <eos> in $\hat{y}_{1:t}$
19:     end for
20: end for
```

updates of the parameters of the policy (line 14), critic (line 15), and average entropy (line 17) are performed. Actor and critic each use a separate learning rate schedule (α_A and α_C).

Figure 1 visualizes the interaction of the BIP-NMT system with a human for a single translation: Feedback is requested when the model is uncertain or the translation is completed. It is directly used for a model update and, in case it was good, for filling the prefix buffer, before the model moves to generating the next (longer) partial translation.

4 Experiments

We simulate a scenario where the learning NMT system requests online bandit feedback for partial translations from a human in the loop. The following experiments will give an initial practical assessment of our proposed interactive learning algorithm. Our analysis of the interactions between actor, critic and simulated human will provide further insights into the learning behavior of BIP-NMT.

4.1 Setup

Data and Preprocessing. We conduct experiments on French-to-English translation on Europarl (EP) and News Commentary (NC) domains. The large EP parallel corpus is used to pre-train the actor in a fully-supervised setting with a standard maximum likelihood estimation objective. The critic network is not pre-trained. For interactive training with bandit feedback, we extract 10k sentences from the NC corpus. Validation and test sets are also chosen from the NC domain. Note that in principle more sentences could be used, however, we would like to simulate a realistic scenario where human feedback is costly to obtain. Data sets were tokenized and cleaned using Moses tools (Koehn et al., 2007). Furthermore, sentences longer than 50 tokens were removed from the training data. Each language's vocabulary contains the 50K most frequent tokens extracted from the two training sets. Table 1 summarizes the data statistics.

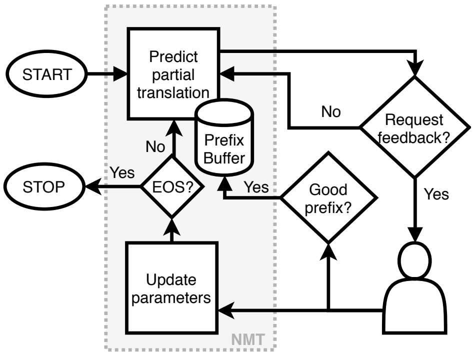

Figure 1: Interaction of the NMT system with the human during learning for a single translation.

Dataset	EP (v.5)	$\bar{n}$	NC (WMT07)	$\bar{n}$
Training (filt.)	1,346,679	23.5	9,216	21.9
Validation	2,000	29.4	1,064	24.1
Test	-	-	2,007	24.8

Table 1: Number of parallel sentences and average number of words per sentence in target language (en), denoted by $\bar{n}$, for training (filtered to a maximum length of 50), validation and test sets for French-to-English translation for Europarl (EP) and News Commentary (NC) domains.

Model Configuration and Training. Following Nguyen et al. (2017), we employ an architecture of two independent but similar encoder-decoder frameworks for actor and critic, respectively, each using global-attention (Luong et al., 2015) and unidirectional single-layer LSTMs[1]. Both the size of word embedding and LSTM's hidden cells are 500. We used the Adam Optimizer (Kingma and

[1] Our code can be accessed via the link `https://github.com/heidelkin/BIPNMT`.

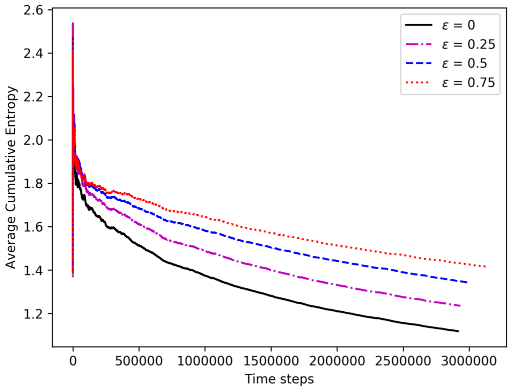

Figure 2: Average cumulative entropy during one epoch of BIP-NMT training with $\mu = 0.8$ and $\epsilon = \{0, 0.25, 0.5, 0.75\}$.

Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. During supervised pre-training, we train with mini-batches of size 64, and set Adam's $\alpha = 10^{-3}$. A decay factor of 0.5 is applied to α, starting from the fifth pass, when perplexity on the validation set increases. During interactive training with bandit feedback, we perform true online updates (i.e. mini-batch size is 1) with Adam's α hyper-parameter kept constant at 10^{-5} for both the actor and the critic. In addition, we clip the Euclidean norm of gradients to 5 in all training cases.

Baselines and Evaluation. Our supervised out-of-domain baseline consists of the actor NMT system described as above, pre-trained on Europarl, with optimal hyperparameters chosen according to corpus-level BLEU on the validation set. Starting from this pre-trained EP-domain model, we further train a bandit learning baseline by employing Nguyen's (2017) actor-critic model, trained on one epoch of sentence-level simulated feedback. The choice of comparing models after one epoch of training is a realistic simulation of a human-system interaction on a sequence of data where each input is seen only once. The feedback signal is simulated with chrF, using character-n-grams of length 6 and a value of $\beta = 2$ of the importance factor of recall over precision. While during training exploration through sampling is essential, during inference and for final model evaluation we use greedy decoding. We evaluate the trained models on our test set from the NC-domain using average sentence-level chrF and standard corpus-level BLEU (Papineni et al., 2002) to measure how well they got adapted to the new domain.

4.2 Results and Analysis

Table 2 shows the results of an evaluation of a baseline NMT model pre-trained by maximum likelihood on out-of-domain data. This is compared to an actor-critic baseline that trains the model of Nguyen et al. (2017) on sentence-level in-domain bandit feedback for one epoch. This approach can already improve chrF (+0.95) and BLEU (+0.55) significantly by seeing bandit feedback on in-domain data. BIP-NMT, with optimal hyperparameters $\epsilon = 0.75$, $\mu = 0.8$ chosen on the validation set, is trained in a similar way for one epoch, however, with the difference that even weaker sub-sentence level bandit feedback is provided on average 5 times per input. We see that BIP-NMT significantly improves both BLEU (+2.18) and chrF (+2.04) by even larger margins.

Table 3 analyzes the impact of the metaparameter ϵ of the BIP-NMT algorithm. We run each experiment three times and report mean results and standard deviation. ϵ controls the margin by which the average word-level entropy needs to increase with respect to the running average in order to trigger a feedback request. Increasing this margin from 0 to 0.25, 0.5 and 0.75 corresponds to decreasing the number of feedback requests by a factor of 3 from around 16 to around 5. This reduction corresponds to a small increase in chrF (+0.29) and a small decrease in BLEU (-0.47).

Figure 2 shows another effect of the metaparameter ϵ: It shows the variation of the average word-level entropy $\bar{H}$ over time steps of the algorithm during one epoch of training. This is computed as a cumulative average, i.e., the value of $\bar{H}$ is accumulated and averaged over the number of target tokens produced for all inputs seen so far. We see that average cumulative entropy increases in the beginning of the training, but then decreases rapidly, with faster rates for smaller values of ϵ, corresponding to more updates per input.

The metaparameter μ controls the threshold of the reward value that triggers a reuse of the prefix for forced decoding. In our experiments, we set this parameter to a value of 0.8 in order to avoid re-translations of already validated prefixes, even if they might sometimes lead to better final full translations. We found the effect of lowering μ from 1.0 to 0.8 negligible on the number of feedback requests and on translation quality but beneficial for the usability.

System	chrF (std)	BLEU (std)	Δ chrF	Δ BLEU
Out-of-domain NMT	61.30	24.77	0	0
Nguyen et al. (2017)	62.25 (0.08)	25.32 (0.02)	+0.95	+0.55
BIP-NMT (ϵ = 0.75, μ = 0.8)	63.34 (0.12)	26.95 (0.12)	+2.04	+2.18

Table 2: Evaluation of pre-trained out-of-domain baseline model, actor-critic learning on one epoch of sentence-level in-domain bandit feedback (Nguyen et al., 2017) and BIP-NMT with settings $\epsilon = 0.75$, $\mu = 0.8$ trained on one epoch of sub-sentence level in-domain bandit feedback. Results are given on the NC test set according to average sentence-level chrF and corpus-level BLEU. Result differences between all pairs of systems are statistically significant according to `multeval` (Clark et al., 2011).

ϵ	chrF (std)	BLEU (std)	Avg # Requests	Δ chrF	Δ BLEU	Δ Avg # Requests
0	61.86 (0.06)	25.54 (0.17)	15.91 (0.01)	0	0	0
0.25	62.15 (0.17)	25.84 (0.13)	11.06 (0.07)	+0.29	+0.3	-5
0.5	61.95 (0.05)	25.46 (0.09)	7.26 (0.03)	+0.09	-0.08	-9
0.75	62.15 (0.04)	25.07 (0.12)	4.94 (0.02)	+0.29	-0.47	-11

Table 3: Impact of entropy margin ϵ on average sentence-level chrF score, corpus BLEU and average number of feedback requests per sentence on the NC validation set. The feedback quality threshold μ is set to 0.8 for all models.

4.3 Example Protocols

Table 4 presents user-interaction protocols for three examples encountered during training of BIP-NMT with $\epsilon = 0.75, \mu = 0.8$. For illustrative purposes, we chose examples that differ with respect to the number of feedback requests, the use of the prefix buffer, and the feedback values. Prefixes that receive a feedback $\geq \mu$ and are thus stored in the buffer and re-used for later samples are indicated by underlines. Advantage scores < 0 indicate a discouragement of individual tokens and are highlighted in red.

In the first example, the model makes frequent feedback requests (in 8 of 17 decoding steps) and fills the prefix buffer due to the high quality of the samples. The second example can use the prefix buffer only for the first two tokens since the feedback varies quite a bit for subsequent partial translations. Note how the token-based critic encourages a few phrases of the translations, but discourages others. The final example shows a translation where the model is very certain and hence requests feedback only after the first and last token (minimum number of feedback requests). The critic correctly identifies problematic parts of the translation regarding the choice of prepositions.

5 Conclusion

We presented a novel algorithm, coined BIP-NMT, for bandit interactive-predictive NMT using reinforcement learning techniques. Our algorithm builds on advantage actor-critic learning (Mnih et al., 2016; Nguyen et al., 2017) for an interactive translation process with a human in the loop. The advantage over previously presented algorithms for interactive-predictive NMT is the low human effort for producing feedback (a translation quality judgment instead of a correction of a translatioin), even further reduced by an active learning strategy to request feedback only for situations where the actor is uncertain.

We showcased the success of BIP-NMT with simulated feedback, with the aim of moving to real human feedback in future work. Before deploying this algorithm in the wild, suitable interfaces for giving real-valued feedback have to be explored to create a pleasant user experience. Furthermore, in order to increase the level of human control, a combination with the standard paradigm that allows user edits might be considered in future work.

Finally, our algorithm is in principle not limited to the application of NMT, but can furthermore — thanks to the broad adoption of neural sequence-to-sequence learning in NLP — be extended to other structured prediction or sequence generation tasks.

Acknowledgments.

This work was supported in part by DFG Research Grant RI 2221/4-1. We would like to thank the members of the Statistical NLP Colloquium Heidelberg for fruitful discussions and ideas for improvement of our algorithm.

SRC depuis 2003 , la chine est devenue le plus important partenaire commercial du mexique après les etats-unis .	
REF since 2003 , china has become mexico ’s most important trading partner after the united states . < /s>	
Partial sampled translation	**Feedback**
since	1
since 2003 , china has	1
since 2003 , china has become	1
since 2003 , china has become mexico	1
since 2003 , china has become mexico ’s	1
since 2003 , china has become mexico ’s most	1
since 2003 , china has become mexico ’s most important	1
since 2003 , china has become mexico ’s most important trading partner after the us . < /s>	0.8823

SRC la réponse que nous , en tant qu’ individus , acceptons est que nous sommes libres parce que nous nous gouvernons nous-mêmes en commun plutôt que d’ être dirigés par une organisation qui n’ a nul besoin de tenir compte de notre existence .	
REF the answer that we as individuals accept is that we are free because we rule ourselves in common , rather than being ruled by some agency that need not take account of us . < /s>	
Partial sampled translation	**Feedback**
the	1
the answer	1
the answer we	0.6964
the answer we ,	0.6246
the answer we as individuals allow to 14 are	0.6008
the answer we , as individuals , go down to speak 8 , are being free because we govern ourselves , rather from being based together	0.5155
the answer we , as people , accepts is that we principle are free because we govern ourselves , rather than being led by a organisation which has absolutely no need to take our standards . < /s>	0.5722

SRC lors d’ un rallye “journée jérusalem” tenu à l’ université de téhéran en décembre 2001 , il a prononcé l’ une des menaces les plus sinistres du régime .	
REF at a jerusalem day rally at tehran university in december 2001 , he uttered one of the regime ’s most sinister threats . < /s>	
Partial sampled translation	**Feedback**
in	0
in a round of jerusalem called a academic university in teheran in december 2001 , he declared one in the most recent hostility to the regime . < /s>	0.5903

Table 4: Interaction protocol for three translations. These translations were sampled from the model when the algorithm decided to request human feedback (line 10 in Algorithm 1). Tokens that get an overall negative reward (in combination with the critic), are marked in red, the remaining tokens receive a positive reward. When a prefix is good (i.e. $\geq \mu$, here $\mu = 0.8$) it is stored in the buffer and used for forced decoding for later samples (underlined).

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA.

Bahdanau, Dzmitry, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France.

Barrachina, Sergio, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and Juan-Miguel Vilar. 2008. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Bentivogli, Luisa, Nicola Bertoldi, Mauro Cettolo, Marcello Federico, Matteo Negri, and Marco Turchi. 2016. On the evaluation of adaptive machine translation for human post-editing. *IEEE Transactions on Audio, Speech and Language Processing (TASLP))*, 24(2):388–399.

Chen, Jinying, Andrew Schein, Lyle Ungar, and Martha Palmer. 2006. An empirical study of the behavior of active learning for word sense disambiguation. In *Human Language Technologies: The 2006 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, New York City, NY.

Clark, Jonathan, Chris Dyer, Alon Lavie, and Noah

Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, OR.

Domingo, Miguel, Álvaro Peris, and Francisco Casacuberta. 2018. Segment-based interactive-predictive machine translation. *Machine Translation*.

Foster, George, Pierre Isabelle, and Pierre Plamondon. 1997. Target-text mediated interactive machine translation. *Machine Translation*, 12(1-2):175–194.

González-Rubio, Jesús, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2011. An active learning scenario for interactive machine translation. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI)*, Barcelona, Spain.

González-Rubio, Jesús, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, France.

Green, Spence, Sida I. Wang, Jason Chuang, Jeffrey Heer, Sebastian Schuster, and Christopher D. Manning. 2014. Human effort and machine learnability in computer aided translation. In *Proceedings the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar.

Judah, Kshitij, Saikat Roy, Alan Fern, and Thomas G. Dietterich. 2010. Reinforcement learning via practice and critique advice. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, Atlanta, GA.

Karimova, Sariya, Patrick Simianer, and Stefan Riezler. 2017. A user-study on online adaptation of neural machine translation to human post-edits. *CoRR*, abs/1712.04853.

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

Knowles, Rebecca and Philipp Koehn. 2016. Neural interactive translation prediction. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, Austin, TX.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Birch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic.

Konda, Vijay R. and John N. Tsitsiklis. 2000. Actor-critic algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada.

Kreutzer, Julia, Artem Sokolov, and Stefan Riezler. 2017. Bandit structured prediction for neural sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada.

Lewis, David D and William A Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Dublin, Ireland.

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

Marie, Benjamin and Aurélien Max. 2015. Touch-based pre-post-editing of machine translation output. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

Mnih, Volodymyr, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, New York, NY.

Nguyen, Khanh, Hal Daumé, and Jordan Boyd-Graber. 2017. Reinforcement learning for bandit neural machine translation with simulated feedback. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.

Ortiz-Martínez, Daniel, Ismael García-Varea, and Francisco Casacuberta. 2010. Online learning for interactive statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Los Angeles, CA.

Papineni, Kishore, 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.

Popović, Maja. 2015. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translat ion (WMT)*, Lisbon, Portugal.

Sokolov, Artem, Julia Kreutzer, Christopher Lo, and Stefan Riezler. 2016. Stochastic structured prediction under bandit feedback. In *Advances in Neural Information Processing Systems (NIPS)*, Barcelona, Spain.

Sutton, Richard S. and Andrew G. Barto. 2017. *Reinforcement Learning. An Introduction*. The MIT Press, second edition.

Sutton, Richard S., David McAllester, Satinder Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processings Systems (NIPS)*, Vancouver, Canada.

Tang, Min, Xiaoqiang Luo, and Salim Roukos. 2002. Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Pennsylvania, PA.

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

Wuebker, Joern, Spence Green, John DeNero, Sasa Hasan, and Minh-Thang Luong. 2016. Models and inference for prefix-constrained machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany.