Interactive-Predictive Neural Machine Translation through Reinforcement and Imitation

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

Abstract

We propose an interactive-predictive neural machine translation framework for easier model personalization using reinforcement and imitation learning. During the interactive translation process, the user is asked for feedback on uncertain locations identified by the system. Responses are weak feedback in the form of "keep" and "delete" edits, and expert demonstrations in the form of "substitute" edits. Conditioning on the collected feedback, the system creates alternative translations via constrained beam search. In simulation experiments on two language pairs our systems get close to the performance of supervised training with much less human effort.

1 Introduction

Despite recent success reports on neural machine translation (NMT) reaching human parity (Wu et al., 2016; Hassan et al., 2018), professional use cases of NMT require model personalization where the NMT system is adapted to user feedback provided for suggested NMT outputs (Wuebker et al., 2018; Michel and Neubig, 2018). In this paper, we will focus on the paradigm of interactive-predictive machine translation (Foster et al., 1997; Barrachina et al., 2008) which has been shown to fit easily into the sequence-to-sequence model of NMT (Knowles and Koehn, 2016; Wuebker et al., 2016). The standard interactive-predictive protocol takes a human-corrected prefix as conditioning context in predicting a sentence completion, which is again corrected or accepted by the human user. Recent work showed in simulation experiments that human effort can be reduced by asking humans for reward signals or validations of partial system outputs instead of for corrections (Lam et al., 2018; Domingo et al., 2017).

Our goal is to combine both feedback modes — corrections and rewards — by treating them as expert demonstrations and reward values in an interactive protocol that combines imitation learning (IL) (Ross et al., 2011) and reinforcement learning (RL) (Sutton and Barto, 2018), respectively, using only limited human edits. A further difference of our framework to standard interactive-predictive NMT is our use of an uncertainty criterion that reduces the amount of feedback requests to the tokens where the entropy of the policy distribution is highest. This idea has been used successfully before in Lam et al. (2018) and Peris and Casacuberta (2018) and connects our work to the area of active learning (Settles and Craven, 2008). Lastly, our framework differs from prior work by allowing model updates based on partial translations.

Our experiments show that weak feedback in form of keep/delete rewards on translation outputs yields consistent improvements of between 2.6 and 4.3 BLEU points over the pre-trained baseline. On one language pair, it even matches the improvements gained by forcing word substitutions from reference translations into the re-decoded output. Furthermore, both feedback scenarios considerably reduce human effort.

2 Related Work

Interactive-predictive translation goes back to early approaches for IBM-type (Foster et al., 1997; Foster et al., 2002) 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. However, in their work the system parameters are not updated based on the prefix. This idea is implemented in Turchi et al. (2017), Michel and Neubig (2018), Wuebker et al. (2018), Karimova et al. (2018), or Peris et al. (2017). In contrast to our work, these approaches use complete post-edited sentences to update their system, while we update our model based on partial translations. Furthermore, our approach employs techniques to reduce the number of interactions.

Our work is also closely related to approaches for interactive pre-post-editing (Marie and Max, 2015; Domingo et al., 2017). The core idea is to ask the translator to mark good segments and use these for a more informed re-decoding, while we integrate constraints derived from diverse human feedback to interactively improve decoding. Additionally, we try to reduce human effort by minimizing the number of feedback requests and by frequent model updates.

Several recent approaches to reinforcement learning from human feedback implement the idea of reinforcing/penalizing a targeted set of actions. Kreutzer et al. (2018) presented an approach were ratings from human users on full translations are used successfully for NMT domain adaptation. Simulations of NMT systems interacting with human feedback have been presented firstly by Kreutzer et al. (2017), Nguyen et al. (2017), or Bahdanau et al. (2017), who apply different policy gradient algorithms, William's REINFORCE (Williams, 1992) or advantage-actor-critic methods (Mnih et al., 2016), respectively. In this paper, we use REINFORCE update strategies for simulated bandit feedback on the sub-sentence level.

González-Rubio et al. (2011; 2012) apply active learning for interactive machine translation, where a user interactively finishes translations of a statistical MT system. Their active learning component decides which sentences to sample for translation and receive supervision for, and the MT system is updated on-line (Ortiz-Martínez et al., 2010). In our algorithm, the active learning component decides which prefixes to receive feedback for based on the entropy of the policy distribution.

3 Learning Interactive-Predictive NMT from Rewards and Demonstrations

As shown in Cheng et al. (2018), IL and RL can be viewed as a single algorithm that only differs in the choice of the oracle, based on objective functions that are defined as the expected value function with respect to the current model's policy π_n in case of RL, and as the expected value function with respect to an expert policy π^* in case of IL. Applied to NMT, both IL and RL are based on a Markov Decision Process where a deterministic sequence of states consisting of the source input and the history of the model's predictions (possibly incorporating expert's demonstrations) serves as conditioning context to predict the respective word, or "action" (Bahdanau et al., 2017).

We instantiate rewards and demonstrations to the feedback types in interactive-predictive translation as follows: In the first case, uncertain words predicted by the system receive a positive or negative reward based on "keep" or "delete" feedback respectively. In the second case, uncertain words can additionally be corrected based on an expert policy in the form of "substitute" feedback associated with a positive reward. This feedback is integrated in context of the model's own predictions by adding rules to constrained beam search decoding (Hokamp and Liu, 2017; Post and Vilar, 2018).[1]

3.1 Learning Objective

We formalize the objective of interactive-predictive NMT as maximizing the value function V of a parametrized policy π_θ, i.e., we seek to maximize the expected (future) reward obtainable from interactions of the NMT system with a human translator who, by editing translations, implicitly assigns rewards $R(\hat{\mathbf{y}})$ to system predictions $\hat{\mathbf{y}}$ given source sentences $\mathbf{x}$:

$$\max_\theta V_{\pi_\theta}(\hat{\mathbf{y}};\mathbf{x}) = \max_\theta \mathbb{E}_{\hat{\mathbf{y}}\sim\pi_\theta(\cdot|\mathbf{x})}[R(\hat{\mathbf{y}})] \qquad (1)$$

[1] We observe that the distinction between weak feedback and expert feedback is difficult to make in the "keep" feedback case: on the one hand, this type of feedback refers to an action generated by the system, and on the other hand, it can be seen as a form of expert demonstration. From this perspective, our first system is closer to RL while our second system is closer to IL. For brevity, we will refer to our models as "RL model" and "IL model", respectively.

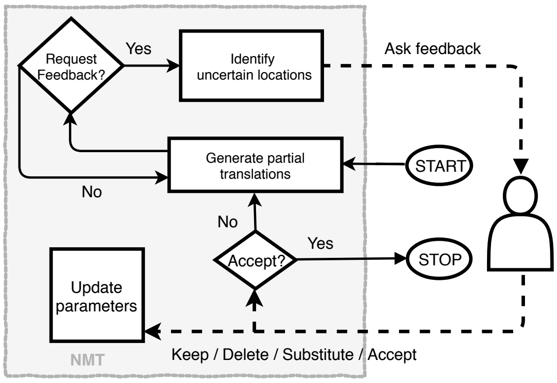
Request Feedback?
Yes
Identify uncertain locations
Ask feedback
Generate partial translations
START
No
No
Yes
Accept?
STOP
Update parameters
NMT
Keep / Delete / Substitute / Accept

Figure 1: A graphical illustration of the interactive-predictive workflow of our system. Dotted arrows indicate interactions between human and system; solid arrows indicate procedures within the system.

Following the policy gradient theorem (Sutton et al., 2000; Bahdanau et al., 2017), its derivative is

$$\nabla_\theta V_{\pi_\theta} = \mathbb{E}_{\hat{\mathbf{y}} \sim \pi_\theta(\cdot|\mathbf{x})} \sum_{t=1}^{T} \sum_{y \in \mathcal{V}} \nabla_\theta \pi_\theta(y|\mathbf{x}, \hat{\mathbf{y}}_{<\mathbf{t}}) R(y) \tag{2}$$

where $\mathcal{V}$ is a vocabulary of target words. In our application, we ask for feedback on a single trajectory at each round of interactions. Similar to Williams (1992), we consider a 1-sample estimate to reduce the inner sum of actions at each time step to the single action $\hat{y}_t$ presented to the user.

Depending on the type of feedback, the instantaneous reward $R(\hat{y}_t)$ for a system translation $\hat{y}_t$ is set to the following values:

$$R(\hat{y}_t) = \begin{cases} 0.5 & \text{if SUBSTITUTE/KEEP}, \\ -0.1 & \text{if DELETE}. \end{cases} \tag{3}$$

In addition, we found that flooring rewards for tokens that do not receive explicit feedback to a small number[2] stabilizes the training and improves performance on the dev set.

4 Algorithms

In this section, we present the details of our interactive-predictive workflow and describe the system components of our implementation to reduce human effort while maintaining high quality model adaptation. In contrast to existing approaches where full sentences are corrected in each round, our system stops decoding when the generated segment meets several (un)certainty criteria. Our system then identifies uncertain words within the generated segment and asks the user to edit these words. The idea is to direct the user to possible translation errors in the segment, and to collect feedback on these highly informative locations, effectively implementing an active learning strategy. The collected feedback is used twice: first, it is used to perform an on-line update of the system's parameters, and secondly, it is integrated as rules into constrained beam search. The full translation is reached after several interactive rounds when the translator finally accepts the translation. Figure 1 gives a graphical illustration of the workflow.

4.1 Measuring uncertainty

We define a measure of uncertainty based on the entropy at a time step t given a set of actions $\mathcal{V}$ (i.e., the target vocabulary) where

$$H_t = -\sum_{y \in \mathcal{V}} \pi_\theta(y|\mathbf{x}, \hat{\mathbf{y}}_{<\mathbf{t}}) \log \pi_\theta(y|\mathbf{x}, \hat{\mathbf{y}}_{<\mathbf{t}}).$$

The idea is that learning from edits on high entropy time steps is more helpful than learning from edits on low entropy time steps, because updating parameters based on uncertain regions better stabilizes the model over time. Furthermore, entropy is computationally simple and far less expensive than external reward estimators such as a quality estimation system, a critic, or a discriminator.

A single token at time step t is considered uncertain if the entropy exceeds a defined threshold ϵ, i.e., $H_t > \epsilon$. We use this criterion to identify informative locations of a partial translation on which the user is asked for feedback.

In case of partial translations, a sequence of length t is considered uncertain if the token at time t is uncertain as defined above, and there is an abrupt change in entropy at t, formally $\frac{H_t - H_{t-1}}{H_{t-1}} > \delta$. Both criteria are applied to determine the length of a partial translation shown to the user.

4.2 Interactive-predictive workflow

Algorithm 1 describes the workflow in our interactive-predictive machine translation scenario. In the first round, the system starts with initial model parameters θ_0, and an empty set of feedback rules ξ, and calls BEAM-SEARCH to first generate an unconstrained partial translation of length t by evaluating the uncertainty criteria in function

[2] We apply Gaussian noise with mean 0.1 and standard deviation of 0.05.

Algorithm 1: Interactive-predictive workflow for a single sentence using constrained beam search. *Input:* model parameters θ, source sentence $\mathbf{x}$, beam size k, learning rate α. *Output:* updated θ^*.

```
$t_{prefix} \leftarrow 1, n \leftarrow 1$
$\theta_0 \leftarrow \theta, \xi \leftarrow \emptyset$
SET-NMT-SOURCE ($\mathbf{x}$)
repeat
    $\hat{\mathbf{y}}_{1:t} \leftarrow$ BEAM-SEARCH ($k$, $t_{prefix}$, $T_{\max}$, $\xi$)
    for $i \leftarrow 1$ to $t$ do
        if UNCERTAIN-LOCATION ($\hat{\mathbf{y}}_{1:t}, i$) then
        Collect feedback rules $\xi_i$
    Get rewards for $\xi_i \in \{keep, delete, substitute\}$
    according to Eq. 3
    $\theta_n \leftarrow \theta_{n-1} + \alpha \nabla_\theta V$ (Eq. 2)
    $t_{prefix} \leftarrow |\hat{\mathbf{y}}_{1:t}|, n \leftarrow n + 1$
until $\hat{\mathbf{y}}_{1:t}$ accepted
```

Algorithm 2: Constrained beam search for uncertain partial translation. *Input:* beam size k, prefix length p, maximum length N, feedback rules ξ. *Output:* partial translation.

```
function BEAM-SEARCH ($k$, $p$, $N$, $\xi$)
    beam $\leftarrow$ DECODER-INIT ($k$)
    for $t \leftarrow 1$ to $N$ do
        scores $\leftarrow$ DECODER-STEP (beam)
        beam $\leftarrow$ KBEST (scores, $k$, $\xi$)
        if LENGTH (beam[0]) $> p$ and
        IS-UNCERTAIN (beam[0]) then break
    return beam[0]
function KBEST (scores, $k$, $\xi$)
    $scores_c \leftarrow$ APPLY-CONSTRAINTS (scores, $\xi$)
    beam $\leftarrow$ ARGMAX$_k$ ($scores_c$)
    return beam
```

IS-UNCERTAIN. The algorithm then evaluates each token within the partial translation and asks for user feedback if the token is considered uncertain w.r.t. the function UNCERTAIN-LOCATION.

Feedback is captured in form of rules that correspond to edits on specific locations, e.g., KEEP token at position i, DELETE token at position i, or SUBSTITUTE token at position i with another token. After collecting the rewards for feedback rules ξ_i according to Equation 3, the model parameters are updated by taking a gradient step as defined in Equation 2.

The updated system then proceeds to the next round by calling BEAM-SEARCH again, this time with a set of feedback rules ξ to generate a constrained partial translation exceeding the previous length t_{prefix}. The uncertainty criterion of tokens is evaluated again and the user is asked for feedback on these tokens, extending the set of feedback rules ξ, which are used to update the system parameters and generate the next partial translation until the user is satisfied with the translation.

4.3 Constrained beam search

A central component is a modified beam search algorithm that takes positional constraints into account (Algorithm 2). The user constraints force the system to generate alternative translations and can thus be interpreted as an exploration strategy. An efficient alternative exploration strategy is multinomial sampling. In our interactive-predictive scenario, however, it is crucial that translations on locations without explicit user feedback are preserved, and this cannot be modeled easily with multinomial sampling. Beam search on the other hand ensures stable translations due to its deterministic nature, and the idea of constrained beam search provides the tools to improve the translation interactively. As a side effect, higher quality translations can be obtained by increasing the beam size at the cost of computational power.

After initializing k beams, the algorithms generates a partial translation by calling DECODER-STEP to retrieve the next token and score all hypotheses. The constraints (provided in the form of feedback rules) are applied in the function KBEST by filtering out all hypotheses that do not satisfy the constraints before the ARGMAX$_k$ operation selects the k highest scoring remaining hypotheses. The single best partial translation is shown to the user only if two conditions are met: (1) the length exceeds the length of the previous partial translation, and (2) the current partial translation is considered an uncertain sequence. In case one condition is not met, the system iteratively extends the partial translation up to a maximum hypothesis length.

5 Experiments

To demonstrate the effectiveness of our reinforcement and imitation strategies, we simulate the interactive-predictive workflow described in Section 4 in a domain adaptation setup. A human translator is simulated by comparing partial translations with corresponding gold translation to extend the set of feedback rules in every round. In the RL setting, the simulated human translator provides only weak feedback (KEEP and DELETE edits) on tokens generated by the system, while in the IL setting the simulated translator addition-

	Data	Training	train / dev / test	∅ en-length
fr-en	EP	pre-training	1.3M / 2k / –	25.5
	NC	interactive	18.4k / 3k / 5k	22.8
de-en	EP	pre-training	1.7M / 2.7k / –	24.0
	NC	interactive	18.9k / 1k / 2k	22.6

Table 1: Data used in pre- and interactive training for French-English (fr-en) and German-English (de-en).

ally injects expert feedback (SUBSTITUTE edit) by demonstrating how the system should act at a specific time step. In our simulation experiments, we focus on the uncertain tokens of the partial translation. An exact match between the uncertain token and the reference generates a KEEP edit, while differing tokens generate either a DELETE or SUBSTITUTE edit depending on the type of system. Tokens exceeding the sentence length of the reference always receive a DELETE feedback. We refer to the first system as KEEP+DELETE, and the second system as +SUBSTITUTE. While the system parameters are updated online after every such simulated interaction, system evaluation is done by a standard offline translation of an unseen test set.

5.1 Dataset

For pre-training, we use the Europarl (EP) corpus version 5 for the French-English system, and version 7 for German-English. For interactive training, we use the News Commentary (NC) 2006 corpus. Both corpora are publicly available on the WMT13's homepage.[3] All experiments are conducted on two language pairs, i.e., German-English (de-en) and French-English (fr-en). Data sets were tokenized and lowercased using MOSES preprocessing scripts (Koehn et al., 2007). We applied compound splitting on the German source sentences using CDEC's tool (Dyer et al., 2010). Our data sets for interactive training differ from the original News Commentary data splits as follows: (1) we sample a subset of the original training set to reduce the number of parallel sentences to 18,432 for French-English and 18,927 for German-English, and (2) we increase both validation and test set for French-English to 3,001 and 5,014 parallel sentences by moving data from the original training set excluding sentences that were sampled for training. Note that a training set size of less than 19,000 parallel sentences is very small even in a domain adaptation setup. Table 1 summarizes the statistics of our datasets.

5.2 Model Architecture

We use a single uni-directional LSTM layer with global attention mechanism between encoder and decoder. The dimensionality of the LSTM hidden states and the word embeddings are 500. We build the vocabulary using the most frequent 50,000 words in each language.

The Adam optimizer (Kingma and Ba, 2014) is used in all training scenarios. In supervised training, we use a mini-batch size of 64 and an initial learning rate of 0.001. Starting from the 5^{th} epoch, the rate is reduced by half in each epoch if the validation perplexity increases. In interactive training, we train for a single epoch and apply a constant learning rate of 10^{-5} with a mini-batch size of 1. In all experiments we set entropy parameters to $\epsilon = 1$, $\delta = 0.5$, and use a beam size of 5 during training. For testing, we apply greedy decoding. PyTorch code of our models is publicly available.[4]

5.3 Results and Discussion

On both language pairs, the optimal pre-trained NMT models are obtained in the 6th training epoch, forming the out-of-domain baseline. We also compare our RL/IL strategies with full post-edits simulated by supervised training on the in-domain News Commentary data, forming an in-domain upper bound. We repeated each experiment three times and report mean and standard deviation for both Character-F[5] (ChrF) (Popović, 2015) and corpus BLEU (Papineni et al., 2002).

In the French-English experiments, both our imitation and reinforcement strategies show improvements of more than 3 points in BLEU and 1 point in ChrF over the out-of-domain baseline. Both strategies achieve lower BLEU score than training on full post-edits, in particular, 0.94 points lower in the KEEP+DELETE setting, and 0.58 points lower in +SUBSTITUTE setting. However, both strategies achieve higher ChrF scores, i.e., 0.76 points for KEEP+DELETE and 0.28 points for +SUBSTITUTE. See upper half of Table 2 for a summary.

In the German-English experiments, there is a bigger performance gap between the KEEP+DELETE and the full post-edits system, concretely, 0.64 points in ChrF score and

[3] https://www.statmt.org/wmt13/

[4] https://github.com/heidelkin/IPNMT_RL_IL

[5] Using parameters $ngram = 6$ and $\beta = 2$.

Pair	System	ChrF (σ)	ΔChrF	BLEU (σ)	ΔBLEU	$\varnothing$ rounds	$\varnothing$ keep+delete / subst.
fr-en	Pre-trained	61.08	–	24.70	–	–	–
	Full Post Edits	61.96 (0.15)	+0.88	29.10 (0.09)	+4.40	–	–
	KEEP+DELETE	**62.72** (0.11)	+1.64	28.16 (0.14)	+3.46	3.2	13.7 / –
	+SUBSTITUTE	62.24 (0.08)	+1.16	**28.52** (0.10)	+3.82	3.3	1.8 / 5.6
de-en	Pre-trained	59.34	–	22.66	–	–	–
	Full Post Edits	60.24 (0.25)	+0.9	27.40 (0.22)	+4.74	–	–
	KEEP+DELETE	59.57 (0.19)	+0.23	25.28 (0.09)	+2.62	3.3	13.1 / –
	+SUBSTITUTE	**60.73** (0.14)	+1.39	**26.91** (0.1)	+4.25	3.3	1.8 / 5.9

Table 2: Character-F (ChrF), and BLEU test results on the French-English (fr-en) and German-English (de-en) translation tasks. Highest scores on RL and IL systems are printed in bold. The Δ columns indicate the score differences to the pre-trained baseline system. All scores are averaged over three runs with standard deviation σ in parentheses.

2.12 points in BLEU lower than full post-edits. However, the improvement over the pre-trained model amounts to 2.62 BLEU points and 0.25 points in ChrF score. Our +SUBSTITUTE system is comparable in performance to the full post-edits system, yielding a result that is 0.49 lower in BLEU but 0.49 points higher in ChrF. See lower half of Table 2 for the summary.

We also report average numbers of feedback rounds and rules per sentence in Table 2. We optimized the maximum number of allowed feedback rules per round on the dev set and use 9 (fr-en) and 7 (de-en) for the KEEP+DELETE and 3 for the +SUBSTITUTE systems. Even for the simpler model based on only weak feedback, the number of user clicks is between 13.7 and 13.1, which is well below the average target sentence length of 22.8 and 22.6. By allowing expert SUBSTITUTE feedback that actively generates better tokens in the next round the number of rules is reduced to 7.4 and 7.7. Our experiments indicate that focusing on uncertain locations can reduce human translation effort substantially.

Effect of on-line learning. We also examine the effect of on-line learning on average cumulative entropy of the model's policy distribution over time. Figure 2 visualizes the change of entropy during interactive training. At the beginning, the system is in regions of high entropy but quickly learns from human edits and the curves become smooth and monotonic. After this initial phase, the overall better performing French-English task shows consistently lower entropy than the German-English task, indicating a connection between model's entropy and translation quality. However, the comparison between the KEEP+DELETE and the better performing +SUBSTITUTE systems shows the opposite trend and requires a different explanation. We conjecture that the +SUBSTITUTE system's expert demonstrations at uncertain locations help the system to find better translations, but such demonstrations also move the system to higher entropy regions, effectively implementing a useful exploration strategy. In contrast to this, the KEEP+DELETE system always stays in more certain regions by selecting another high probability token if the original token receives a DELETE feedback by the user.

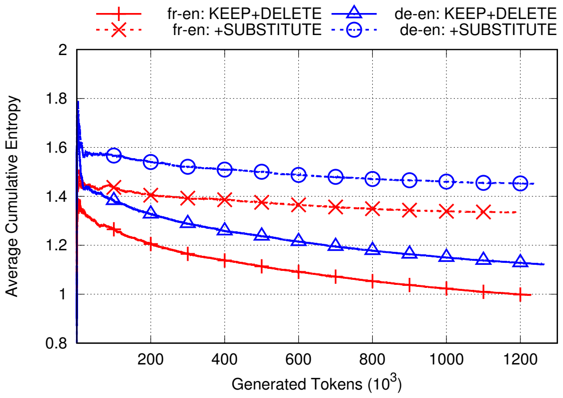

Figure 2: Average cumulative entropy of the model's policy distribution over time during simulated interactive learning. Plots are shown for the French-English (fr-en) and the German-English (de-en) task, and for the KEEP+DELETE and the +SUBSTITUTE system, respectively.

Effect of beam size. The observations on model's entropy over time in the previous paragraph and the implementation details described in Section 4.3 show that our constrained beam search implements exploration in a user-controlled manner. We conjecture that beam size also influences the exploration and should have a different effect on different feedback strategies. We thus conduct additional experiments using beam sizes of 2, 5, 10 and 20 on all language pairs and the two systems. The results are summarized in Figure 3. In both KEEP+DELETE and +SUBSTITUTE systems, a beam size of 2 is sufficient to achieve substantial gains over the baselines in both language pairs. In case of the KEEP+DELETE system, increasing beam sizes only marginally influence the translation performance. In case of the +SUBSTITUTE system, there are considerable gains of almost 1 BLEU point and 1 Character-F point when increasing the beam size from 2 to 5. Here, the larger beam size enables the system to connect the expert demonstrations with better prefixes which helps the system to explore higher scoring trajectories. Increasing the beam size to 10 or 20 further improves performance but the gains are small.

Decoding Speed. The total runtime of each of our simulated interactive experiments is roughly 6 hours when simulated on a Nvidia P40, while training of the KEEP+DELETE system is slightly slower than of the +SUBSTITUTE system due to the higher number of feedback rules. Looking at the sentence level this means the total decoding time of our system for all partial translations of a single sentence is $6 \times 1\text{h}/(18,432 \times 3.3) = 0.361\text{s}$ for the French-English task, and even less for the German-English task. This estimate does not account for the time our system conducts validation tests or constructs simulated feedback, thus the actual average processing time is lower. Knowles and Koehn (2016) argue that beam search is usually too slow to be used for training in interactive live systems, however, recent hardware developments together with our strategy of partial decoding makes constrained beam search applicable even in training. As a side effect, corrections on early time steps reduce the problem of error propagation and thus improve both usability of the system and satisfaction of the translator.

Leveraging BPE or character-level NMT. Our current implementation of interactive-predictive NMT uses a word-based translation approach and presents word units to users for feedback. An adaptation of our algorithm to sub-word or character level NMT is possible and requires to redistribute the reward associated to the word level to sub-word units or characters, and to maintain their location information in the constrained beam search. We leave this extension to future work.

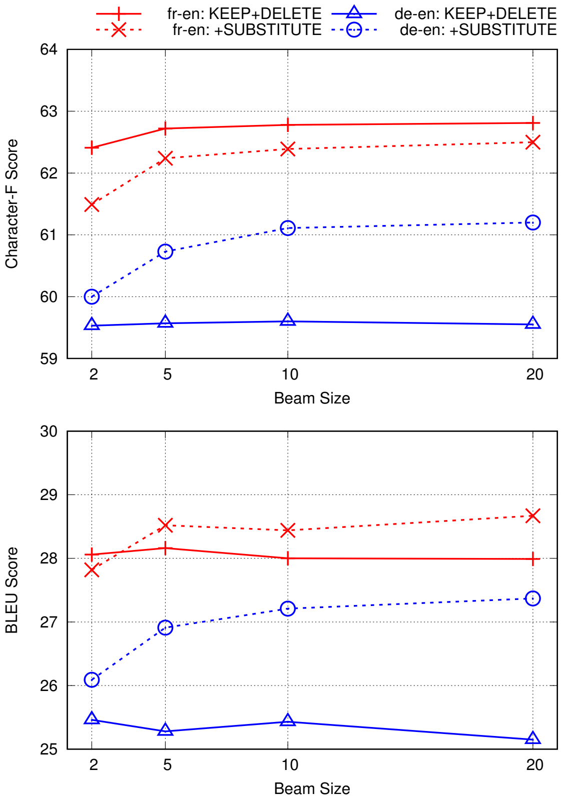

Figure 3: The two figures show the effect of different beam sizes on Character-F score (top) and BLEU score (bottom). We conduct experiments on French-English (fr-en) and German-English (de-en) and both systems (KEEP+DELETE and +SUBSTITUTE). All scores are averaged over two runs.

5.4 Examples

Table 3 illustrates the translation workflow of our interactive-predictive protocol by listing four examples: the upper half shows example translations of the two systems for the German-English task, the lower half shows two examples of the systems for the French-English task.

The first example is taken from the KEEP+DELETE system, where our simulated user provides only KEEP and DELETE feedback on suggested locations. In interactive round 1 on the German-English task, the system stops after generating the uncertain partial translation "the core" and asks the user for feedback specifically on the term "core". The simulated user returns a DELETE feedback and the system is able to generate the more appropriate translation "heart of the problem" in round 2. In round 3, however, a weakness of the simulated feedback becomes apparent: our user gives a negative DELETE feedback on the token "amount" because the token differs from the given reference word "quantity", even though it is an appropriate translation for the German word "Menge" in this context. The system then generates "volume" in round 4 and "supply" in the final round 5, although both translations are worse than the initially proposed translation "amount". One explanation for this behavior is the way on-line updates are applied to the NMT system: while the constrained beam search implements feedback rules on token level, the on-line updates of the NMT system take place on the word embedding level. An update based on negative feedback actually forces the NMT system to avoid semantically similar words. In the above example, the negative feedback for "amount" downgrades the optimal translation "quantity" because of the semantic similarity of both words, and instead upgrades the more diverse translations "volume" and "supply". In our example, this strategy has an immediate negative impact on translation quality, but it also illustrates the positive exploration effect which is helpful in the long run.

The second example is taken from the +SUBSTITUTE system, where the simulated user additionally provides "substitute" feedback. In interactive round 1, the system generates the uncertain partial translation "the south koreans are" and identifies "the" and "are" as uncertain tokens. The user suggests to change "the" to "as", and "are" to "south" by providing SUBSTITUTE feedback. Again, a limitation of our simulation becomes apparent: our simulated substitutions are based on reference translations, but a real translator would not change the given partial translation to "as south korean south". Still, based on the two feedback rules and the on-line update, the NMT system is able to follow a better trajectory in round 2. We observe that SUBSTITUTE feedback is a very strong signal that supports the system to quickly get close to the translation our simulated user has in mind (which is the reference in our simulation).

The French-English task examples illustrate a noteworthy property of our algorithm: In round 3 of the KEEP+DELETE system, the simulated user provides DELETE feedback on the tokens "to hate their" only because they occur at different positions compared to the reference. However, the system is able to recover and re-generate the tokens at the correct position in round 5. A similar behavior can be observed for the +SUBSTITUTE system in round 3, where the phrase "bring about macroeconomic" is first substituted and then generated again in the final round 4.

6 Conclusion

In this work, we integrate interactive-predictive NMT with imitation learning and reinforcement learning. Our goal is to merge the human edit process with effort reduction and model learning into a single framework for easier model personalization. Our results indicate that on-line learning from edits on uncertain locations of partial translations can achieve performance comparable to using supervised learning on in-domain data but with substantially less human effort. In the future, we would like to investigate the limitations of entropy-based uncertainty measures, work on the efficiency of the training speed, and conduct field studies with human users.

Acknowledgments.

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

References

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.

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

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

Dyer, Chris, Jonathan Weese, Hendra Setiawan, Adam Lopez, Ferhan Ture, Vladimir Eidelman, Juri Ganitkevitch, Phil Blunsom, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations (ACL Demo)*, Uppsala, Sweden.

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

Foster, George, Philippe Langlais, and Guy Lapalme. 2002. User-friendly text prediction for translators. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, PA.

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 of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Hokamp, Chris and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *ACL*, Vancouver, Canada.

Karimova, Sariya, 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.

Kingma, Diederik P and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Knowles, Rebecca and Philipp Koehn. 2016. Neural interactive translation prediction. In *North American component of the International Association for Machine Translation (AMTA)*, Austin, TX, USA.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions (ACL Demo)*, Prague, Czech Republic.

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.

Kreutzer, Julia, 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)*.

Lam, Tsz Kin, 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.

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.

Michel, Paul and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

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 of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA.

Peris, Álvaro 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.

Peris, Álvaro, Luis Cebrián, and Francisco Casacuberta. 2017. Online learning for neural machine translation post-editing. *CoRR*, abs/1706.03196.

Popović, Maja. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal.

Post, Matt and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, New Orleans, LA, USA.

Ross, Stéphane, Geoffrey J. Gordon, and J. Andrew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics (AISTATS)*, Fort Lauderdale, FL, USA.

Settles, Burr 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.

Sutton, Richard S. and Andrew G. Barto. 2018. *Reinforcement Learning. An Introduction*. The MIT Press, second edition.

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

Turchi, Marco, 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 (PBML)*, 108(1):233–244, jun.

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

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

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.

Wuebker, Joern, Patrick Simianer, and John DeNero. 2018. Compact personalized models for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium.

German-English	**Source**	der kern des problems ist nicht die gesamt‿menge des öls , sondern seine lage .
	Reference	the heart of the problem is not the overall quantity of oil , but its location .
	Round	**Partial translation** → **FEEDBACK**
	1	the core$_2$
		→ DELETE(2)
	2	the *heart*$_2$ of the problem is not the total$_9$
		→ KEEP(2), → DELETE(9)
	3	the *heart* of the problem is not the *overall* amount$_{10}$ of oil , but its$_{15}$
		→ DELETE(10), → KEEP(15)
	4	the *heart* of the problem is not the *overall* volume$_{10}$ of oil , but *its* situation$_{16}$.
		→ DELETE(10, 16)
	5	the *heart* of the problem is not the *overall supply* of oil , but *its position* .
		→ accepted.
German-English	**Source**	die süd‿koreaner ihrerseits verlassen sich darauf , dass china mit der nuklearen krise in nord‿korea fertig wird .
	Reference	as for the south koreans , they are counting on china to deal with the north korean nuclear crisis .
	Round	**Partial translation** → **FEEDBACK**
	1	the$_1$ south koreans are$_4$
		→ SUBSTITUTE(1:as, 4:south)
	2	*as* for the$_3$ *south* koreans , china$_7$
		→ KEEP(3:the), → SUBSTITUTE(7:they)
	3	*as* for *the south* koreans , *they* are relying$_9$ on china to be$_{13}$
		→ SUBSTITUTE(9:counting, 13:deal)
	4	*as* for *the south* koreans , *they* are *counting* on china to *deal* with the nuclear crisis in north korea .
		→ accepted.
French-English	**Source**	il est dur d' aimer ou de respecter un peuple et de haïr son état .
	Reference	it is hard to love or respect a people and hate their state .
	Round	**Partial translation** → **FEEDBACK**
	1	it is hard to love$_5$
		→ KEEP(5)
	2	it is hard to *love* or to$_7$
		→ DELETE(7)
	3	it is hard to *love* or *comply*$_7$ with a people and to$_{12}$ hate$_{13}$ their$_{14}$
		→ DELETE(7, 12, 13, 14)
	4	it is hard to *love* or *respect*$_7$ a$_8$ people and hatred$_{11}$.$_{12}$
		→ KEEP(7, 8),→ DELETE(11, 12).
	5	it is hard to *love* or *respect a* people and *to hate their state* .
		→ accepted.
French-English	**Source**	un gouvernement qui n' est pas en mesure d' équilibrer ses propres finances ne peut pas apporter une stabilité macroéconomique .
	Reference	a government that cannot balance its own finances cannot be relied on to provide macroeconomic stability .
	Round	**Partial translation** → **FEEDBACK**
	1	a government that is$_4$
		→ SUBSTITUTE(4:cannot)
	2	a government that *cannot* balance its own$_7$
		→ KEEP(7)
	3	a government that *cannot* balance its *own* finances cannot bring$_{10}$ about$_{11}$ macro-economic$_{12}$ stability .
		→ SUBSTITUTE(10:be,11:relied,12:on)
	4	a government that *cannot* balance its *own* finances cannot *be relied on* to bring about macro-economic stability .
		→ accepted.

Table 3: Interaction protocol illustrating translation progress of the two learning systems on the German English task (upper half) and French-English (lower half). For each language pair, the first example illustrates interactions with the KEEP+DELETE system, while the second example shows interactions with the +SUBSTITUTE system. In each round, the user is asked for feedback on uncertain locations of the current partial translation. Tokens printed in blue with their position in subscript indicate uncertain locations. At the end of each round, the system is updated given the user's feedback (KEEP, DELETE, SUBSTITUTE). In the next round, it generates a constrained (partial) translation with respect to this feedback. Tokens generated based on feedback rules are printed in *italics*.