Sentiment Preservation in Review Translation using Curriculum-based Re-inforcement Framework

Divya Kumari*, Soumya Chennabasavraj**, Nikesh Garera** and Asif Ekbal*
*Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India
**Flipkart, India
*divya_1921cs10,asif@iitp.ac.in
**soumya.cb,nikesh.garera@flipkart.com

Abstract

Machine Translation (MT) is a common approach to feed humans or machines in a cross-lingual context. However, there are some expected drawbacks. Studies suggest that in the cross-lingual context, MT system often fails to preserve different stylistic and pragmatic properties of the source text (e.g. sentiment, emotion, gender traits, etc.) to the target translation. These disadvantages can degrade the performance of any downstream Natural Language Processing (NLP) applications, such as sentiment analyser, that heavily relies on the output of the MT systems (especially in a low-resource setting). The susceptibility to sentiment polarity loss becomes even more severe when an MT system is employed for translating a source content that lacks a legitimate language structure (e.g. review text). Therefore, while improving the general quality of the Neural Machine Translation (NMT) output (e.g. adequacy), we must also find ways to minimize the sentiment loss in translation. In our current work, we present a deep re-inforcement learning (RL) framework in conjunction with the curriculum learning to fine-tune the parameters of a full-fledged NMT system so that the generated translation successfully encodes the underlying sentiment of the source without compromising the adequacy, unlike the previous method. We evaluate our proposed method on the English–Hindi (product domain) and French–English (restaurant domain) review datasets, and found that our method (*further*) brings a significant improvement over a full-fledged supervised baseline for the machine translation and sentiment classification tasks.

1 Introduction

Product and/or service reviews available in the e-commerce portals are predominantly in the English language, and hence a large number of population can not understand these. Machine Translation (MT) system can play a crucial role in bridging this gap by translating the user-generated contents, and directly displaying them, or making these available for the downstream Natural Language Processing (NLP) tasks e.g. sentiment classification[1] (Araújo et al., 2020; Barnes et al., 2016; Mohammad et al., 2016; Kanayama et al., 2004). However, numerous studies (Poncelas et al., 2020a; Afli et al., 2017; Mohammad et al., 2016; Sennrich et al., 2016a) have found a significant loss of sentiment during the automatic translation of the source text.

The susceptibility to sentiment loss aggravates when the MT system is translating a noisy review that lacks a legitimate language structure at the origin. For example, a noisy review contains several peculiarities and informal structures, such as shortening of words (e.g. "awesome" as "awsm"), acronyms (e.g. "Oh My God" as "OMG"), phonetic substitution of numbers (e.g. "before" as "b4"), emphasis on characters to define extremity of the emotion (e.g. "good" as "gooooooood"), spelling mistakes, etc. Unfortunately, even a pervasively used commercial neural

[1] Please note that our current work is limited to cross-lingual sentiment analysis [CLSA] via MT based approach.

machine translation (NMT) system, Google Translate, is very brittle and easily falters when presented with such noisy text, as illustrated through the following example.
Review Text (English): I found an awsome product. (Positive)
Google Transliteration (Hindi): mujhe ek ajeeb utpaad mila. (Neutral)
The example shows how the misspelling of a sentiment bearing word "awesome" gets this positive expression translated to a neutral expression. In the above context, if an unedited raw MT output is directly fed to the downstream sentiment classifier, it might not get the expected classification accuracy. Thus, in this work we propose a deep-reinforcement-based framework to adapt the parameters of a pre-trained neural MT system such that the generated translation improves the performance of a cross-lingual multi-class sentiment classifier (without compromising the adequacy).

More specifically, we propose a deep *actor-critic (AC)* reinforcement learning framework in the ambit of *curriculum learning* (CL) to alleviate the issue of sentiment loss while improving the quality of translation in a cross-lingual setup. The idea of actor-critic is to have two neural networks, *viz.* (i). an actor (i.e. a pre-trained NMT) that takes an action (policy-based), and (ii). a critic that observes how good the action taken is and provides feedback (value-based). This feedback acts as a guiding signal to train the actor. Further, to better utilize the data, we also integrate curriculum learning into our framework.

Recently, Tebbifakhr et al. (2019) demonstrated that an MT system (actor) can be customised to produce a controlled translation that essentially improves the performance of a cross-lingual (binary) sentiment classifier. They achieved this task-specific customisation of a "generic-MT" system via a policy-based method that optimizes a task-specific metric, i.e. $F1$ score (see Section 2). However, this often miserably fails to encode the semantics of the source sentence.

Recent studies (Xu et al., 2018) demonstrated that the non-opinionated semantic content improves the quality of a sentiment classifier. Accordingly, the transfer of such information from the source to the target can be pivotal for the quality of the sentiment classifier in a cross-lingual setup. Towards this, we investigate the optimization of a harmonic-score-based reward function in our proposed RL-based framework that ensures to preserve both sentiment and semantics. This function operates by taking a weighted harmonic mean of two rewards: (i). content preservation score measured through Sentence-level BLEU or SBLEU; and (ii). sentiment preservation score measured through a function that performs element-wise dot product between a predicted sentiment distribution and the gold sentiment distribution.

Empirical results, unlike Tebbifakhr et al. (2019), suggest that our RL-based fine-tuning framework, tailored to optimize the harmonic reward, preserves both sentiment and semantics in a given NMT context. Additionally, we also found that the above fine-tuning method in the ambit of curriculum learning achieves an additional performance gain of the MT system over a setting where curriculum based fine-tuning is not employed. The core of *curriculum learning* (CL) (Bengio et al., 2009) is to design a metric that scores the difficulty of training samples, which is then used to guide the order of presentation of samples to the learner (NMT) in an easy-to-hard fashion. To the best of our knowledge, this is the very first work that studies the curriculum learning (CL) for NMT from a new perspective, i.e. given a pre-trained MT model, the dataset to fine-tune, and the two tasks *viz.* sentiment and content preservation; we utilize a reward-based metric (i.e. harmonic score) to define the difficulty of the tasks and use it to score the data points. The use of harmonic reward based scoring/ranking function implicitly covers the tasks' overall difficulty through a single metric.

Moreover, understanding that obtaining a gold-standard polarity annotated data is costlier, the fine-tuning of pre-trained NMT model is performed by re-using only a small subset of the supervised training samples that we annotated with respect to (w.r.t) their sentiment. Empirical results suggest that additionally enriching a random small subset of the training data with extra sentiment information, and later re-using them for the fine-tuning of the referenced model via our proposed framework (c.f. Section 3) observes an additional gain in BLEU and $F1$ score over a supervised baseline. We summarize the main contributions and/or the key attributes of our current work as follows:
(i). create a new domain-specific (i.e. product review) parallel corpus, a subset of which is annotated for their sentiment;
(ii). propose an AC-based fine-tuning framework that utilizes a novel harmonic mean-based reward function to meet our two-fold objectives, *viz.* enabling our NMT model to preserve; (a).

the non-opinionated semantic content; and (b). the source sentiment during translation.
(iii). Additionally, we utilize the idea of CL during the RL fine-tuning of the pre-trained model and try to learn from easy to hard data, where hard corresponds to the instances with lower harmonic reward. To the best of our knowledge, this is the first work in NMT that studies CL in the ambit of RL fine-tuning.

2 Related Work

The use of translation-based solution for cross-lingual sentiment classification is successfully leveraged in the literature (Wu et al., 2021; Tebbifakhr et al., 2020; Araújo et al., 2020; Poncelas et al., 2020b; Fei and Li, 2020; Tebbifakhr et al., 2019; Akhtar et al., 2018; Barnes et al., 2016; Balahur and Turchi, 2012; Kanayama et al., 2004) which suggest an inspiring use-case of the MT system, and brings motivation for this piece of work.

Given the context of this work, we look at the pieces of works that address the preservation of sentiment in the automatic translation. In one of the early works, Chen and Zhu (2014) used a lexicon-based consistency approach to design a list of sentiment-based features and used it to rank the candidates of *t*-table in a Phrase based MT system. Lohar et al. (2017, 2018) prepared the positive, negative and neutral sentiment-specific translation systems to ensure the cross-lingual sentiment consistency.

Recently, Tebbifakhr et al. (2019) proposed Machine-Oriented (MO) Reinforce, a policy-based method to pursue a machine-oriented objective[2] in a sentiment classification task unlike the traditional human-oriented objective [3]. It gives a new perspective for a use-case of the MT system (i.e. machine translation for machine). To perform this task-specific adaption (i.e. produce output to feed a machine), Tebbifakhr et al. (2019) adapted the REINFORCE of Williams (1992) by incorporating an exploration-oriented sampling strategy. As opposed to one sampling of REINFORCE, MO Reinforce samples k times, $(k = 5)$, and obtains a reward for each sample from the sentiment classifier. A final update of the model parameters are done w.r.t the highest rewarding sample. Although they achieved a performance boost in the sentiment classification task, they had to greatly compromise with the translation quality. In contrast to Tebbifakhr et al. (2019), we focus on performing a task-specific customisation of a pre-trained MT system via a harmonic reward based deep reinforcement framework that uses an AC method in conjunction with the CL. The adapted NMT system, thus obtained, is expected to produce a more accurate (high-quality) translation as well as improve the performance of a sentiment analyser. Bahdanau et al. (2017); Nguyen et al. (2017), unlike us, used the popular AC method, and focused only on preserving the semantics (translation quality) of a text. Additionally, we develop a CL based strategy to guide the training. Recently, Zhao et al. (2020) also studied AC method in the context of NMT. However, they used this method to learn the curriculum for re-selecting influential data samples from the existing training set that can further improve the performance (translation quality) of a pre-trained NMT system.

3 Methodology

Firstly, we perform the pre-training of a NMT model until the convergence using the standard log-likelihood (LL) training on the supervised dataset (c.f. Table 1: (A)). The model, thus obtained, acts as our referenced MT system/actor. To demonstrate the improvements brought by the proposed curriculum-based AC fine-tuning over the above LL-based baseline in the sentiment preservation and machine translation tasks, we carry out the task-specific adaption of the pre-trained LL-based MT model (actor) by re-using a subset of the supervised training samples. It is worth mentioning here that, in the fine-tuning stage, the actor does not observe any new sentence, rather re-visit (randomly) a few of the supervised training samples which are now additionally annotated with their sentiment (c.f. Section 4).

Actor-critic Overview : Here, we present a brief overview of our AC framework which is discussed at length in the subsequent section. In the AC training, the actor (NMT) receives an input sequence, s, and produces a sample translation, $\hat{t}$, which is evaluated by the critic model.

[2]Where the MT objective is to feed a machine.

[3]Where the MT objective is to feed the human.

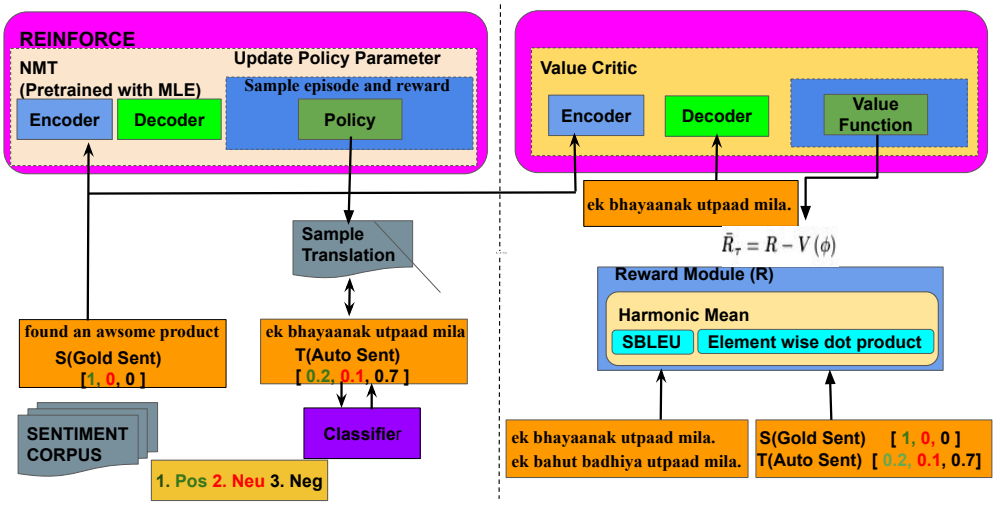

Figure 1: An illustration of the Actor-Critic Method

The critic feedback is used by the actor to identify those actions that bring it a better than the average reward. In the above context, a feedback of a random critic would be useless for training the actor. Hence, similar to the actor we warm up the critic for one epoch by feeding it samples from the pre-trained actor, while the actor's parameters are frozen. We then fine-tune these models jointly so that - as the actor gets better w.r.t its action, the critic gets better at giving feedback (see Section 4.2 for the dataset and reward used in the pre-training and fine-tuning stages). The details of the loss functions that the actor and critic minimizes are discussed in Section 3.1.

Furthermore, to better utilize the data, we finally integrate CL into our AC framework (our proposed approach). Empirical results (Section 5.1) show that during fine-tuning, presenting the data in an easy-to-hard fashion yields a better learned actor model over the one obtained via *vanilla* (no-curriculum based) fine-tuning. Our proposed framework brought improvements over several baselines without using any additional new training data in the two translation tasks, i.e. (i). English–Hindi[4] and (ii). French–English [5]. Since our proposed framework is a combination of RL via AC method and CL, we first present the details of the main components of the AC model alongside their training procedure in Section 3.1. The details of the reward model are presented in Section 3.2, and then introduce the plausibility of CL in Section 3.3. Finally, we describe our proposed CL-based AC framework in Algorithm 1.

3.1 Proposed Fine-tuning Method

The architecture of our AC-based framework is illustrated in Figure 1. It has three main components *viz.* (i). an actor : the pre-trained neural agent (NMT) whose parameters define the policy and the agent takes action, i.e. sample translations according to the policy (ii). a reward model : a score function used to evaluate the policy. It provides the actual (true) estimated reward to the translations sampled from the model's policy. To ensure the preservation of sentiment and content in translation, the chosen reward model gives two constituent rewards - a classifier-based score and a SBLEU score (Section 3.2), respectively, and (iii). a critic : a deep neural function approximator that predicts an expected value (reward) for the sampled action. This is then used to center the true reward (step (ii)) from the environment (see Equation 2). Subtracting critic estimated reward from the true reward helps the actor to identify action that yields extra reward beyond the expected return. We employ a critic with the same architecture as of the actor.

We see from the lower-left side of Figure 1 that, for each input sentence (s), we draw a

[4]Trained using a new parallel dataset created as a part of this work, a subset of which is polarity annotated.

[5]Trained using publicly available dataset, a part of it is additionally annotated with sentiment.

single sample ($\hat{t}$) from the actor, which is used for both estimating gradients of the actor and the critic model as explained in the subsequent section.

Critic Network training: During the RL training, we feed a batch of source sentences, $B_j\,(s)$, to the critic encoder and the corresponding sampled translations obtained from the actor, $B_j\,(\hat{t})$, to the decoder of the critic model. The critic decoder then predicts the rewards (i.e. value estimates, V_ϕ, predicted for each time step of the decoder), and accordingly updates its parameters supervised by the actual (or true) rewards, $\mathrm{R}(\hat{t}, s)$ [6] (steps to obtain this reward is discussed in Section 3.2) from the environment.

The objective of the critic network is, thus, to find its parameter value ϕ that minimizes the *mean square error* (MSE) between the *true reward* (see R in Figure 1) from the environment, and the critic *estimated reward* (i.e. values predicted by the critic, see V_ϕ in Figure 1). Accordingly, the MSE loss that the critic minimizes is as in Equation (1), where $\tau^{'}$ being the critic decoding step.

$$\nabla_\phi \mathrm{L}_{crt}(\phi) \approx \sum_{\tau^{'}=1}^{n} \left[\mathrm{V}(\hat{t}_{<\tau^{'}}, s) - \mathrm{R}(\hat{t}, s)\right] \nabla_\phi \mathrm{V} \tag{1}$$

Note that in this work we explore the setting, where the reward, R, is observable only at time step $\tau = n$ of the actor (a scalar for each complete sentence). Thus, to calculate the difference terms in Equation 1 for n steps, we use the same terminal reward, R, in all the intermediate time steps of the critic decoder.

Actor Network training: To update the actor (G) parameters, θ, we use the policy gradient loss; weighted by a reward which is centered via the critic estimated value (i.e. the critic estimated value, V, is subtracted from the true reward, R, from the environment), as in equation (2). The updated reward is finally used to weigh the policy gradient loss, as shown in (3), where τ being the decoding step of the actor.

$$\bar{\mathrm{R}}_\tau(\hat{t}, s) = \mathrm{R}(\hat{t}, s) - \mathrm{V}(\hat{t}_{<\tau^{'}}, s) \tag{2}$$

$$\nabla_\theta \mathrm{L}^{\mathrm{pg}}_{\mathrm{actor}}\,(\theta) \approx \sum_{\tau=1}^{n} \bar{\mathrm{R}}_\tau \nabla_\theta \log \mathrm{G}_\theta(\hat{t}_\tau | \hat{t}_{<\tau}) \tag{3}$$

The actor and the critic both are global-attention based recurrent neural networks (RNN). Algorithm 1 summarizes the overall update framework. We run this algorithm for mini-batches.

3.2 Defining Rewards

As our primary goal is to optimize the performance of the pre-trained NMT system towards sentiment classification and machine translation tasks, accordingly we investigate the utility of the following three reward functions (i.e. true reward, R in Equation 1 as R_1, R_2, R_3) for optimization through our *vanilla* AC method. Please note, for brevity we only choose the reward that serves the best to our purpose (i.e. harmonic reward as it ensures both, an improved cross-lingual sentiment projection, and a high quality translation with our *vanilla* AC approach, as discussed in Section 5.1) for our subsequently proposed curriculum-based experiment. The three types of feedbacks we explored are: (i). Sentence-level BLEU as a reward to ensure the content preservation, also referred as R_1, is calculated following the Equation (4)

$$\mathrm{R}_1 = \mathrm{SBLEU}(\hat{t}, t) \tag{4}$$

(ii). Element-wise dot product between the gold sentiment distribution and predicted sentiment distribution (e.g. $[1, 0, 0]$ and $[0.2, 0.1, 0.7]$ in Figure 1 evaluates to scalar value 0.2) taken from the softmax layer of the target language classifier to ensure sentiment preservation, also referred

[6] Although shown like this, it only means true reward corresponding to a given source sentence and the corresponding sampled action, not as a function.

as R_2. To simulate the target language classifier, we fine-tune the pre-trained BERT model (Devlin et al., 2019). The tuned classifier (preparation steps discussed in Section 4.1) is used to obtain the reward R_2 as in Equation (5).

$$\mathrm{R}_2 = \mathrm{P}(s)_{gold} \bullet \mathrm{P}(\hat{t})_{bert} \tag{5}$$

and,
(iii). Harmonic mean of (i) and (ii) as a reward, also referred to as R_3 to ensure the preservation of both sentiment and semantic during the translation, as in Equation (6).

$$\mathrm{R}_3 = (1+\beta^2)\frac{(2 \cdot \mathrm{R}_1 \cdot \mathrm{R}_2)}{(\beta^2 \cdot \mathrm{R}_1) + \mathrm{R}_2} \tag{6}$$

where β is the harmonic weight which is set to 0.5.

3.3 Curriculum Construction

The core of CL is (i). to design an evaluation metric for difficulty, and (ii). to provide the model with easy samples first before the hard ones.

In this work, the notion of difficulty is derived from the harmonic reward, R_3, as follows. Let, $X = \{x_i\}_{i=1}^{N} = \left\{(s^i, t^i)\right\}_{i=1}^{N}$ denotes the RL training data points. To measure the difficulty of say, i^{th} data point, (s^i, t^i), we calculate the reward, R_3 using $(\hat{t}^i, s^i)$. In order to obtain the corresponding sample translation, $\hat{t}^i$, we use the LL-based model (pre-trained actor). We do this for the N data points. Finally, we sort the RL training data points from easy, i.e., with high harmonic reward, to hard as recorded on their translations. In the fine-tuning step, the entire sorted training data points are divided into mini-batches, $B = [B_1, ..., B_M]$, and the actor processes a mini-batch sequentially from B. Hence, at the start of each epoch of training, the actor will learn from the easiest examples first followed by the hard examples in a sequential manner until all the M batches exhaust. Another alternative is the use of pacing function $f_{pace}(s)$, which helps to decide the fraction of training data available for sampling at a given time step s, i.e. $f_{pace}(s)|D_{train}|$. However, we leave it to explore in our future work. The Pseudo-code for the proposed CL-based AC framework including pre-training is described by Algorithm 1.

4 Datasets and Experimental Setup

In this section, we first discuss the datasets used in different stages of experiments followed by the steps involved in the creation of datasets, the baselines used for comparison, and the model implementation details.

Dataset: Our NMT adaptation experiments are performed across two language pairs from different families with different typological properties, i.e. English to Hindi (henceforth, En–Hi) and French–English (henceforth, Fr–En). We use the following supervised datasets for the pre-training and validation of LL-based NMT in En–Hi and Fr–En tasks,
(i). For En–Hi task, we use a newly created domain-specific parallel corpus (see section 4.1) whose sources were selected from an e-commerce site. This corpus is released as a part of this work. Statistics of the dataset is shown in Table 1 : (A), row(ii).
(ii). For Fr–En task, we concatenate a recently released domain-specific parallel corpus, namely Foursquare (4SQ) corpus [7] (Berard et al., 2019) with first $60K$ sentences from $OpenSubtitles$ [8] corpus to simulate a low-resource setting. The basic statistics of this dataset are shown in Table 1 : (A), row(i). For RL fine-tuning of the LL-based NMT(s), we use the corresponding RL datasets from Table 1: (B). In each task, the RL trainset sentences are a subset of human translated sentences drawn from the supervised training samples and additionally annotated with respect to sentiment. For En–Hi task, these sentences are randomly sampled from the supervised training corpus (c.f. Table 1: (A), row(ii)), and for Fr–En we use 4SQ-HT dataset (c.f. Table 1:

[7] A small parallel restaurant reviews dataset released as a part of the review translation task.

[8] We choose this dataset as it is made of spoken-language sentences which are noisy, sentiment-rich and is closest to 4SQ corpus as suggested by the author.

Algorithm 1 Proposed algorithm (Curriculum based fine-tuning process). In the *vanilla* (i.e. no curriculum-based) approach, we skip steps 5 to 7.

1: Initialize the actor model G_θ with uniform weights $\theta \in [-0.1, 0.1]$.
2: Pre-train the actor (G_θ) with LL loss until convergence.
3: Initialize the critic model V_ϕ with uniform weights $\phi \in [-0.1, 0.1]$.
4: Pre-train the critic for one epoch with SBLEU as a reward on the same LL training data by feeding it samples from the pre-trained actor, while the actor's parameters are frozen.
5: Use the actor model to translate all the data points in X.
6: Obtain rewards R_3 corresponding to N data points.
7: Rank $\{X_i\}_{i=1}^{N} = \left\{(s^i, t^i)\right\}_{i=1}^{N}$ based on R_3.
8: **for** K epochs **do**
9: **for** mini-batches, $B = [B_1, ..., B_M]$ **do**
10: Obtain the sample translations $B_m(\hat{t})$ from the actor for the given source sentences, $B_m(s)$.
11: Obtain R_1, R_2 and finally observe the rewards R_3.
12: Feed the source sentences, $B_m(s)$ to the critic encoder and sampled translations, $B_m(\hat{t})$ to the decoder.
13: Obtain the predicted rewards, V_ϕ, using the critic model.
14: Update the critic's parameter using (1).
15: Obtain final reward $\bar{\mathrm{R}}$ using (2).
16: Update the actor's parameter using (2) in (3).
17: **end for**
18: **end for**

(A), row(i)). To evaluate the performance of all the NMT system(s) we use the corresponding RL testsets from Table 1.

4.1 Data Creation

To the best of our knowledge, there is no existing (freely available) sentiment annotated parallel data for English–Hindi in the review domain. Hence, we crawl the electronic products reviews from a popular e-commerce portal (i.e. Flipkart). These reviews were translated into Hindi using the Google Translate API. A significant part of this automated translation is then verified by a human translator with a post-graduate qualification and proficient in English and Hindi language skills. One more translator was asked to verify the translation. The detailed statistics of new in-domain parallel corpus are shown in Table 1: (A), row(ii). Further, a subset of human translated product reviews is selected randomly for sentiment annotation. Three annotators who are bilingual and experts in both Hindi and English took part in the annotation task. Details of the instructions given to the annotators and translators are mentioned below.

Instructions to the Translators:

For translation, following were the instructions: (i). experts were asked to read the Hindi sentence carefully; (ii). source and target sentences should carry the same semantic and syntactic structure; (iii). they were instructed to carefully read the translated sentences, and see whether the fluency (grammatical correctness) and adequacy are preserved; (iv). they made the correction in the sentences, if required; (v). vocabulary selection at Hindi (target) side should be user friendly; (vi). transliteration of an English can also be used, especially if this is a named entity (NE).

Instructions to the Annotators:

The annotators have post-graduate qualification in linguistics, possessing good knowledge of English and Hindi both. They have prior experience in judging the quality of machine translation and sentiment annotation.

For sentiment labeling, annotators were asked to follow the guidelines as below:

(i). they were instructed to look into the sentiment class of the source sentence (Tebbifakhr et al., 2019) (English), locate its sentiment bearing tokens; (ii). they were asked to observe both of these properties in the translated sentences; (iii). they were asked to annotate the source sentences into the four classes, namely

positive, negative, neutral and others.

The further detailed instructions for sentiment annotation are given as below:
(i). Select the option that best captures the sentiment being conveyed in the sentences:- positive- negative- neutral- others- (ii). Select positive if the sentence shows a positive attitude (possibly toward an object or event). e.g. great performance and value for money. (iii). Select negative if the sentence shows a negative attitude (possibly toward an object or event). e.g. please do not exchange your phone on flipkart they fool you . (iv.) Select neutral if the sentence shows a neutral attitude (possibly toward an object or event) or is an objective sentence. Objective sentences are sentences that do not carry any opinion, e.g. facts are objective expressions about entities, events and their properties. e.g. (a). the selfie camera is 32 mp .(objective), (b). after doing research on the latest phones, i bought this phone . (neutral). (iv) Select others for sentences that do not fall in above three categories, e.g. (a). if the sentence is highly ungrammatical and hard to understand. (b). if the sentence expresses both positive and negative sentiment, i.e. mixed polarity.

These annotation guidelines were decided after thorough discussions among ourselves. After we had drafted our initial guidelines, the annotators were asked to perform the verification of the translated sentences, and sentiment annotation for the 100 sentences. The disagreement cases were thereafter discussed and resolved through discussions among the experts and annotators. Finally, we came up with the set of instructions as discussed above to minimize the number of disagreement cases. Class-wise statistics of the sentiment-annotated dataset for En–Hi task are shown in Table 1: (B). Additionally, the same annotators also annotated a part of the 4SQ corpus (i.e. target [9] (English) sentences from the 4SQ-HT training and 4SQ-test set) to obtain the sentiment annotated RL dataset for the Fr–En task (c.f. Table 1: (B)). For sentiment classification, the inter-annotator agreement ratio (Fleiss, 1971) is 0.72 for En–Hi, and 0.76 for Fr–En. We manually filtered the RL datasets to only include the positive, negative and neutral sentences as per the the manual annotations. We refer these sentiment-annotated corpora as the RL dataset(s).

Classifier training: In order to build the target language sentiment analyser, we use the BERT-based [10] language model. The classifier is first pre-trained using the target-side sentences of the supervised training corpus. Classifier pre-training is followed by the task-specific fine-tuning using the target-side sentences of the RL training set. For example, to build the target language English classifier for the Fr–En task, the classifier is first pre-trained using the English sentences from the supervised dataset (c.f. Table 1: (A), row(i)) followed by fine-tuning by using polarity-labelled English sentences from the RL training corpus (c.f. Table 1: (B), row(i)).

Baselines: Other than the supervised baseline, we also compare our CL-based AC fine-tuning framework with the following state-of-the-art RL-based fine-tuning frameworks, i.e. (1). *REINFORCE*, and (2). *Machine-Oriented Reinforce*. Additionally, we also conduct the ablation study to better analyse the utility of harmonic reward in the task through our *vanilla* AC method as follows: (3). $\mathbf{MT^{ac}_{bert}}$: AC fine-tuning with sentiment reward only; (4). $\mathbf{MT^{ac}_{bleu}}$: AC fine-tuning with content reward only; (5). $\mathbf{MT^{ac}_{har}}$: AC fine-tuning with both the rewards. Finally, for brevity we choose the best performing AC-reward model for the proposed curriculum-based learning.

4.2 Hyper-parameters Setting

In all our experiments, we use an NMT system based on Luong et al. (2015), using a single layer bi-directional RNN for the encoder. All the encoder-decoder parameters are uniformly initialized in the range of [-0.1,0.1]. The sizes of embedding and hidden layers are set to 256 and 512, respectively. The Adam optimizer (Abdalla and Hirst, 2017) with $\beta 1 = 0.9, \beta 2 = 0.99$ is used and the gradient vector is clipped to magnitude 5. We set the dropout to 0.2 and use the input feeding with learning rate (lr) and batch size (bs) set to $1e-3$ and 64. We first perform supervised pre-training of the NMT using the parallel corpora from Table 1: (A), and select the best model parameters according to the perplexity on the development set (c.f. Table 1: (A)). We refer the actor- thus obtained- as $\mathbf{MT_{LL}}$, that acts as a trained policy in the RL training (refer to the upper left side of Figure 1). Then, we keep the actor fixed and warm-up the critic for one epoch with SBLEU reward on the supervised training samples (c.f. Table 1: (A)) with the lr

[9] This is different from En–Hi task setting where we annotate source sentence. We do so due to resource constraint.
[10] We used BERT-Base multilingual uncased model for Hindi and monolingual uncased BERT for English language.

(A)

Task(s)	Corpus		#Sentences
(i). Fr–En	60K-OPUS	(training)	60,000
	4SQ-PE	(training)	12,080
	4SQ-HT	(training)	2,784
	4SQ-valid	(validation)	1,243
(ii). En–Hi		(training)	75,821
		(validation)	700
	Vocabulary	(En-Hi)	(22,031-27,229)
	Avg. Length	(En-Hi)	(16.04-17.30)

(B)

Task(s)	RL trainset			RL devset			RL testset		
	Pos	Neu	Neg	Pos	Neu	Neg	Pos	Neu	Neg
(i). Fr–En	1,469	1,049	241	1,469	1,049	241	870	769	184
(ii). En–Hi	1,147	1,147	1,147	1,147	1,147	1,147	800	800	800

(C)

Task(s)	Metrics	$\mathbf{M_{LL}}$	$\mathbf{M^{mo}_{bert}}$	$\mathbf{M^{r}_{bleu}}$	$\mathbf{M^{ac}_{bert}}$	$\mathbf{M^{ac}_{bleu}}$	$\mathbf{M^{ac}_{har}}$
(i). Fr–En	BLEU	25.02	24.99	25.15	25.04	25.14	**25.18**
	$F1$ score	75.31	75.33	75.65	75.43	75.35	75.39
(ii). En–Hi	BLEU	27.87	28.01	27.75	27.97	28.14	28.13
	$F1$ score	73.14	73.42	73.12	73.12	72.77	73.29

(D)

Models	Fr–En		En–Hi	
	BLEU	$\mathbf{F_1}$ score	BLEU	$\mathbf{F_1}$ score
$\mathbf{M_{LL}}$	25.02	75.31	27.87	73.14
$\mathbf{M^{mo}_{bert}}$	$24.99_{(-0.03)}$	$75.33_{(+0.02)}$	$28.01_{(+0.14)}$	$73.42_{(+0.28)}$
$\mathbf{M^{r}_{bleu}}$	$25.15_{(+0.13)}$	$75.65_{(+0.34)}$	$27.75_{(-0.12)}$	$73.12_{(-0.02)}$
$\mathbf{M^{ac}_{har}}$	$25.18^{*}_{(+0.16)}$	$75.39^{**}_{(+0.08)}$	$28.13^{*}_{(+0.26)}$	$73.29^{*}_{(+0.15)}$
$\mathbf{M^{ac}_{har}}$+CL	$\mathbf{25.26}^{*}_{(+0.24)}$	$75.38^{**}_{(+0.07)}$	$\mathbf{28.18}^{*}_{(+0.31)}$	$73.22^{*}_{(+0.08)}$

Table 1: (A). Supervised parallel corpora for LL training. **(B).** Class-wise distribution of the polarity-tagged RL dataset(s) used to fine-tune the LL-based (En–Hi and Fr–En) pre-trained NMT(s). For the Fr–En task, we annotate a part of the 4SQ corpora (i.e. training: 4SQ-HT and testing: 4SQ-test). We do not keep a separate development set for the fine-tuning of the LL model. **(C).** Results of the fine-tuned *vanilla* reinforcement-based NMT(s). Here, superscripts **mo**, **r** and **ac** refers to the Machine-Oriented Reinforce, REINFORCE and actor-critic approach, respectively to fine-tune the LL-based model(s) ($\mathbf{M_{LL}}$, column (iii); En–Hi: row (i), Fr–En: row (ii)), and the subscripts **bleu, bert** and **har** refers to the corresponding rewards (i.e. SBLEU (R_1), classifier (R_2), and harmonic mean (R_3)) optimized via the policy gradient method. **(D).** Results of curriculum-based fine-tuning and other baselines. Proposed approach is $\mathbf{M^{ac}_{har}}$+CL. * significant at $p < .05$, ** significant at $p < .01$

of $1e-3$ and bs of 64. We employ the same encoder-decoder configuration as of the actor for the critic. In the RL training, we jointly train the actor and the critic model with lr of:- $1e-6$ (Fr–En); $1e-5$ (En–Hi), respectively and bs of 4 on the RL datasets with harmonic reward. For sampling the candidate translation in the RL training, we use multinomial sampling, with a sampling length of 50 tokens. We run the experiments three times with different seed values, and record the $F1$ and BLEU scores (c.f. Table 1: (C)) to evaluate the performance of the sentiment analysers and customised MT systems on the RL testset and report the average of the runs in Section 5. For all the RL-based models, the fine-tuning steps maximize the chosen average reward discussed in Section 3.2 on the RL devset. The fine-tuning continues for a maximum of 20 epochs (including the baselines). The best epoch is chosen based on the performance observed on the RL devset (i.e. the best average rewarding epoch). All the sentences are tokenized. As an extra pre-processing step, we lowercase all the English, French, and normalize all the Hindi sentences. Tokens in the training sets are segmented into sub-word units using the Byte-Pair Encoding (BPE) technique (Sennrich et al., 2016b) with 4,000 merge operations.

For evaluating our customised MT systems over all the baselines, we use the relevant RL testsets.

5 Results and Analysis

We first present the results of fine-tuning the pre-trained MT through different RL-based methods, i.e. (i). REINFORCE (ii). MO Reinforce, and (iii). *vanilla* AC (ours) in Section 5.1. Further, to better analyse the utility of harmonic reward (R_3) in sentiment and content preservation task over the previously studied rewards (i.e. SBLEU: R_1 or BERT: R_2) in the context of NMT (Tebbifakhr et al., 2020, 2019; Nguyen et al., 2017; Bahdanau et al., 2017; Wu et al., 2018; Ranzato et al., 2016), we additionally present the fine-tuning results of the *vanilla* AC method with the following two types of rewards: (i). only content, i.e. R_1 and (ii). only sentiment, i.e. R_2 as a reward.

We choose the best performing reward model (i.e. R_3) among the AC-based NMT(s). At last, we discuss the results in the context of our curriculum-based AC framework. To evaluate the translation quality we record the BLEU score of the RL testset when translated from the relevant models. To validate our claim that the translations obtained by our proposed MT system can further improve the performance of the sentiment classifier in a cross-lingual setup over the baselines, we do the following. We apply the target language sentiment classifier to the translations obtained by the LL-based NMT system vs. all the customised RL-based NMT systems, and record their F_1 scores.

5.1 Evaluation Results

As shown in Table 1: (C), the full-fledged LL-based NMT(s) (trained until convergence as observed on the development sets, column (iii).) obtain the following BLEU points (25.02, 27.87) and F_1 scores (75.31, 73.14) for the Fr–En, En–Hi tasks, respectively. We then perform fine-tuning of the pre-trained models through our *vanilla* AC harmonic approach (by re-visiting only a subset of samples from the existing supervised training sets which are now additionally annotated with their sentiment). We see for both Fr–En and En–Hi that our harmonic-reward-based models can obtain a significant performance boost (further) over the pre-trained baselines in both the optimized (targeted) metrics, i.e. BLEU improved to 25.18 ($+0.16$), 28.13 ($+0.26$) and F_1 scores reached to 75.39 ($+0.08$), 73.29 ($+0.15$) in both the language pairs. This is not the case with other reinforcement-based fine-tuned models - MO Reinforce [11] and REINFORCE that optimizes a single reward for which we observe non-optimized reward drop in at least one language-pair. For example, if we consider the MO Reinforce for the Fr–En task, the non-optimized metric - BLEU drops by -0.03 point (despite an improvement of $+0.02$ point in the optimized metric, F_1 score), and for the REINFORCE in En–Hi task both BLEU and F_1 score drop by -0.12 and -0.02 points, respectively. This establishes the efficacy of our reinforcement method. Further, when we see the results form critic-based fine-tuning of the LL model via two commonly used reward routines (R_1, R_2 - column (vi). and (vii).). As expected, we see an improvement in the targeted metric (e.g. for R_1-based model the optimized reward is BLEU. We can see improvement in BLEU). However, to our surprise, we found that the improvement in BLEU score does not have a high correlation with the performance in the sentiment classification task. For example, in the En–Hi task, the critic model with R_1 as a reward (columns (vii).) observed the highest BLEU score (28.14) but the highest F_1 score (73.29) is observed from the R_3 based model (column (viii).). This suggests the effectiveness of the harmonic reward which successfully improves both BLEU and F_1 score over the supervised baselines for both the language pairs. For the sake of brevity, we choose the harmonic model for our curriculum experiment.

When comparing the performance in the context of our proposed curriculum-based AC framework, the results from Table 1:(D) show that our method is better at producing coherent as well as sentiment-rich translation. By comparing row (i). and row (vi)., we can see that in both Fr–En and En–Hi task, merely learning in an easy-to-hard fashion brings the highest improvement in BLEU scores over the supervised baselines, i.e. $+0.24, +0.31$ point for the Fr–En and En–Hi tasks, respectively. F_1 scores are also improved by $+0.07$ and $+0.08$ point, respectively. All these improvement are statistically significant[12]. Furthermore, we also observe

[11] Please note unlike Tebbifakhr et al. (2019) "out-of-domain" MT-adaption approach ours' LL-based MT is trained using in-domain data.

[12] To test significance, we use bootstrap resampling method (Koehn, 2004) for BLEU and student's t-test for sentiment

that the CL-based fine-tuning observes a faster convergence over the *vanilla* approach.

5.2 Error Analysis for English–Hindi task

Although our proposed method outperforms the LL-based baseline in the sentiment classification task, we also observe several failure cases. To investigate this, we observe the sentiment-conflicting cases, i.e. selected those samples from ours' model where there is an observed disagreement between the predicted and the gold sentiment. From these samples, we filter those examples where the source (English) sentences have an explicit presence of the positive or the negative sentiment expression. Unsurprisingly, we found the main reason for sentiment loss was still the low-translation quality. Secondly, to better understand what policy is learned by our-proposed NMT that brings the observed improvement in the sentiment-classifier performance, we investigate those translations where the LL model has a predicted (by the classifier) sentiment-disagreement, whereas ours' shows an agreement, both with the gold sentiment. We present below one such example.

Review Text (Source): satisfy with overall working. (*positive)* || **Transliteration(Ref.)**: kul meelaakar kaam se santusht hoon. (*positive)* || **Transliteration(Auto.)**: kul milaakar kaam ke saath santusht. (*positive)* ($\text{MT}^{\text{ac}}_{\text{har}}$ + CL) || **Transliteration(Auto.):** kul milaakar kaam kar rahe hain . (*neutral)* (MT_{LL}) . We can see that our proposed model indeed learned to translate the sentiment expressions to their preferred variant (positive sentiment bearing expression *satisfy* translated as *santusht*).

6 Conclusion

In this paper, we have proposed a curriculum-based deep re-inforcement learning framework that successfully encodes both the underlying sentiment and semantics of the text during translation. In contrast to the REINFORCE-based frameworks (Williams, 1992; Tebbifakhr et al., 2019) (actor only models), ours is a critic-based approach that helps the actor learns an efficient policy to select the actions, yielding a high return from the critic. Besides, with the support of curriculum learning, it can be more efficient. This is also established (empirically) through the observed additional boost (significant at $p < .05$) in BLEU score over the baselines. Further, we have manually created a domain-specific (product reviews) polarity-labelled balanced bilingual corpus for English–Hindi, that could be a useful resource for research in the similar areas. We shall make the data and our codes available to the community.

7 Acknowledgement

Authors gratefully acknowledge the unrestricted research grant received from the Flipkart Internet Private Limited to carry out the research. Authors thank Muthusamy Chelliah for his continuous feedbacks and suggestions to improve the quality of work; and to Anubhav Tripathee for gold standard parallel corpus creation and translation quality evaluation.

References

Abdalla, M. and Hirst, G. (2017). Cross-lingual sentiment analysis without (good) translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 506–515, Taiwan. Asian Federation of Natural Language Processing.

Afli, H., Maguire, S., and Way, A. (2017). Sentiment translation for low resourced languages: Experiments on Irish general election tweets. In *18th International Conference on Computational Linguistics and Intelligent Text Processing*, Budapest, Hungry.

Akhtar, M. S., Sawant, P., Sen, S., Ekbal, A., and Bhattacharyya, P. (2018). Solving data sparsity for aspect based sentiment analysis using cross-linguality and multi-linguality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 572–582, New Orleans, Louisiana. Association for Computational Linguistics.

classification.

Araújo, M., Pereira, A., and Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, 512:1078–1102.

Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., Courville, A., and Bengio, Y. (2017). An actor-critic algorithm for sequence prediction. In *5th International Conference on Learning Representations*, Toulon, France.

Balahur, A. and Turchi, M. (2012). Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60, Jeju, Korea. Association for Computational Linguistics.

Barnes, J., Lambert, P., and Badia, T. (2016). Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1613–1623.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Berard, A., Calapodescu, I., Dymetman, M., Roux, C., Meunier, J.-L., and Nikoulina, V. (2019). Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 168–176, Hong Kong. Association for Computational Linguistics.

Chen, B. and Zhu, X. (2014). Bilingual sentiment consistency for statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 607–615.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Fei, H. and Li, P. (2020). Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5759–5771, Online. Association for Computational Linguistics.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Kanayama, H., Nasukawa, T., and Watanabe, H. (2004). Deeper sentiment analysis using machine translation technology. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 494–500, Geneva, Switzerland. COLING.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain.

Lohar, P., Afli, H., and Way, A. (2017). Maintaining sentiment polarity in translation of user-generated content. *The Prague Bulletin of Mathematical Linguistics*, 108(1):73–84.

Lohar, P., Afli, H., and Way, A. (2018). Balancing translation quality and sentiment preservation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 81–88, Boston, MA.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal.

Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2016). How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Nguyen, K., Daumé III, H., and Boyd-Graber, J. (2017). Reinforcement learning for bandit neural machine translation with simulated human feedback. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1464–1474, Copenhagen, Denmark.

Poncelas, A., Lohar, P., Hadley, J., and Way, A. (2020a). The impact of indirect machine translation on sentiment classification. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 78–88, Virtual. Association for Machine Translation in the Americas.

Poncelas, A., Lohar, P., Way, A., and Hadley, J. (2020b). The impact of indirect machine translation on sentiment classification. *arXiv preprint arXiv:2008.11257*.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2016). Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations*, San Juan, Puerto Rico.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.

Tebbifakhr, A., Bentivogli, L., Negri, M., and Turchi, M. (2019). Machine translation for machines: the sentiment classification use case. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1368–1374, Hong Kong, China.

Tebbifakhr, A., Negri, M., and Turchi, M. (2020). Automatic translation for multiple nlp tasks: a multi-task approach to machine-oriented nmt adaptation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 235–244.

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

Wu, H., Wang, Z., Qing, F., and Li, S. (2021). Reinforced transformer with cross-lingual distillation for cross-lingual aspect sentiment classification. *Electronics*, 10(3):270.

Wu, L., Tian, F., Qin, T., Lai, J., and Liu, T.-Y. (2018). A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium.

Xu, J., Sun, X., Zeng, Q., Zhang, X., Ren, X., Wang, H., and Li, W. (2018). Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia.

Zhao, M., Wu, H., Niu, D., and Wang, X. (2020). Reinforced curriculum learning on pre-trained neural machine translation models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9652–9659. AAAI Press.