Online Learning Meets Machine Translation Evaluation: Finding the Best Systems with the Least Human Effort

Vânia Mendonça1,2, Ricardo Rei1,2,3, Luísa Coheur1,2, Alberto Sardinha1,2, Ana Lúcia Santos4,5
1 INESC-ID Lisboa, Portugal
2 Instituto Superior Técnico, Universidade de Lisboa, Portugal
3 Unbabel AI, Lisboa, Portugal
4 Centro de Linguística da Universidade de Lisboa, Portugal
5 Faculdade de Letras da Universidade de Lisboa, Portugal
{vania.mendonca, luisa.coheur, jose.alberto.sardinha}@tecnico.ulisboa.pt, ricardo.rei@unbabel.com, als@letras.ulisboa.pt

Abstract

In Machine Translation, assessing the quality of a large amount of automatic translations can be challenging. Automatic metrics are not reliable when it comes to high performing systems. In addition, resorting to human evaluators can be expensive, especially when evaluating multiple systems. To overcome the latter challenge, we propose a novel application of online learning that, given an ensemble of Machine Translation systems, dynamically converges to the best systems, by taking advantage of the human feedback available. Our experiments on WMT’19 datasets show that our online approach quickly converges to the top-3 ranked systems for the language pairs considered, despite the lack of human feedback for many translations.

1 Introduction

In Machine Translation (MT), measuring the quality of a large amount of automatic translations can be a challenge. Automatic metrics like BLEU (Papineni et al., 2002) remain popular due to their fast and free computations. Yet, in the last few years we have seen that, as MT quality improves, automatic metrics become less reliable (Ma et al., 2019; Mathur et al., 2020). For example, in the Conference on Machine Translation (WMT)’19 News Translation shared task, the winning system according to human annotators was not even in the top-5 according to BLEU (Barrault et al., 2019). On the other hand, using human assessments can be expensive, especially when evaluating multiple systems. In a real world scenario, given an arbitrary number of MT systems, one would need to evaluate them individually to find the best systems for a given language pair. However, that requires a considerable effort and there may not be enough human annotators to evaluate all the systems’ translations. For instance, in the aforementioned WMT’19 shared task, many translations from the competing systems did not receive any human assessment.

Given an ensemble of competing, independent MT systems, how can we dynamically find the best ones for a given language pair, while making the most of existing human feedback? To address this question, we present a novel application of online learning to MT: each MT system in the ensemble is assigned to a weight, and the systems’ weights are updated considering human feedback regarding the quality of their translations at each iteration. We use online learning algorithms with theoretical performance guarantees, under the frameworks of prediction with expert advice (Cesa-Bianchi and Lugosi, 2006) and multi-armed bandits (Robbins, 1952; Lai and Robbins, 1985).

We contribute with an online MT ensemble that allows to reduce human effort by immediately incorporating human feedback in order to dynamically converge to the best systems1. Our experiments on WMT’19 News Translation test sets show that our online approaches indeed converge to the shared task’s official top-3 systems (or to a subset of them) in just a few hundred iterations for all the language pairs experimented. Moreover, it does so while coping with the aforementioned lack of human assessments for many translations, through the use of fallback metrics.

2 Online learning frameworks

To provide some background on our proposal, we start by describing the online learning frameworks that we apply in this paper: prediction with expert advice and multi-armed bandits.

A problem of prediction with expert advice can be described as an iterative game between a fore-
caster and the environment, in which the forecaster seeks advice from different sources (experts) in order to provide the best forecast (Cesa-Bianchi and Lugosi, 2006). At each iteration $t$, the forecaster consults the predictions $\hat{p}_{j,t}, j = 1 \ldots J$, made by a set of $J$ weighted experts, in the decision space $\mathcal{D}$. Considering these predictions, the forecaster makes its own prediction, $\hat{p}_{f,t} \in \mathcal{D}$. At the same time, the environment reveals an outcome $y_t$ in the decision space $\mathcal{Y}$ (which may not necessarily be the same as $\mathcal{D}$).

A well-established algorithm to learn the experts’ weights in this framework is Exponentially Weighted Average Forecaster (EWAF) (Cesa-Bianchi and Lugosi, 2006). In EWAF, the prediction made by the forecaster is randomly selected following the probability distribution based on the experts’ weights $\omega_{1,t}, \ldots, \omega_{J,t-1}$:

$$\hat{p}_{f,t} = \frac{\sum_{j=1}^{J} \omega_{j,t-1} p_{j,t}}{\sum_{j=1}^{J} \omega_{j,t-1}}. \quad (1)$$

At the end of each iteration, the forecaster and each of the experts receive a non-negative loss based on the outcome $y_t$ revealed by the environment ($\ell_{f,t}$ and $\ell_{j,t}$, respectively). The weight $\omega_{j,t}$ of each expert $j = 1 \ldots J$ is then updated according to the loss received by each expert, as follows:

$$\omega_{j,t} = \omega_{j,t-1} e^{-\eta \ell_{j,t}} \quad (2)$$

If the parameter $\eta$ is set to $\sqrt{\frac{\log J}{T}}$, it can be shown that the forecaster quickly converges to the performance of the best expert after $T$ iterations (Cesa-Bianchi and Lugosi, 2006).

Prediction with expert advice assumes that both the forecaster and all the experts receive a loss once the environment’s outcome is revealed. However, this assumption may not always hold (i.e., there may not always be an environment’s explicit feedback or a way to obtain the loss for all the experts). Thus, we consider a related class of problems, multi-armed bandits, in which the environment’s outcome is unknown (Robbins, 1952; Lai and Robbins, 1985). In this class of problems, one starts by attempting to estimate the means of the loss distributions for each expert (also known as arm) in the first iterations (the exploration phase), and when the forecaster has a high level of confidence in the estimated values, one may keep choosing the prediction with the smallest estimated loss (the exploitation phase).

A popular online algorithm for adversarial multi-armed bandits is Exponential-weighting for Exploration and Exploitation (EXP3) (Auer et al., 1995). At each iteration $t$, the forecaster’s action is randomly selected according to the probability distribution given by the weights of each arm $j$:

$$\hat{p}_{f,t} = \frac{\omega_j}{\sum_{j'=1}^{J} \omega_{j'}} \quad (3)$$

In this framework, the forecaster is only able to measure the loss of the action it selects at each iteration, but it cannot measure the loss of other possible actions. Thus, only the weight of the arm associated with this action is updated, as follows:

$$\omega_{j,t} = \omega_{j,t-1} e^{-\eta \ell_{j,t}} \quad (4)$$

where $\ell_{j,t} = \frac{\xi_{j,t}}{p_{j,t}}$ and $p_{j,t}$ is the probability of choosing arm $j$ at iteration $t$. By setting $\eta$ to $\sqrt{\frac{2 \log J}{T |\mathcal{A}|}}$ (where $|\mathcal{A}|$ is the number the actions available, and may be the same as the number of arms $J$), it can be shown that the forecaster quickly converges to the performance of the best arm.

Both of these frameworks are relatively under-explored in NLP, despite their potential to converge to the best performing approach available in scenarios where feedback is naturally present. Therefore, we propose to apply them in order to find the best MT models with little human feedback.

### 3 Machine Translation with Online Learning

In this work, we consider the following scenario as the starting point: there is an ensemble composed of an arbitrary number of MT systems; given a segment from a source language corpus, each system outputs a translation in the target language; then, the quality of the translations produced by each of the available systems is assessed by one or more human evaluators with a score reflecting their quality.

We frame this scenario as an online learning problem under two different frameworks: (i) prediction with expert advice (using EWAF as the learning algorithm), and (ii) multi-armed bandits (using EXP3 as the learning algorithm). The decision on whether to use one or another framework in an MT scenario depends on whether there is human feedback available for the translations outputted by all the available systems or only for the final choice of the ensemble of systems.
An overview of the online learning process is shown in Fig. 1, and can be summed up as follows. Each MT system is an expert (or arm) $j = 1 \ldots J$, associated with a weight $\omega_j$ (all the systems start with same weights). At each iteration $t$, a segment $src_t$ is selected from the source language corpus and handed to all the MT systems. Each system outputs a translation $transl_{j,t}$ in the target language, and one of these translations is selected as the forecaster’s action according to the probability distribution given by the systems’ weights (Eq. 1 for EWAF and Eq. 3 for EXP3). The chosen translation $transl_{f,t}$ (when using EXP3) or the translations outputted by all the systems (when using EWAF) receive a human assessment score\(^2\) $score_{j,t}$, from which the loss $\ell_{j,t}$ is derived for the respective MT system. Finally, the weight of the chosen system or the weights of all the systems are updated as a function of the loss received, according to Eq.4 (when using EXP3) and Eq.2 (when using EWAF), respectively (where $\ell_{j,t} = -score_{j,t}$).

### 4 Experimental setup

To validate our proposal, we designed an experiment using data from an MT shared task. The main questions addressed by our experiment are: (i) whether an online learning approach can give a greater weight to the top performing systems for each language pair according to the shared task’s official ranking, and (ii) if so, how quickly (i.e., how many translations need to be assessed by human evaluators in order to find the best system).

Below we detail the datasets used (Section 4.1) and the feedback sources considered (Section 4.2), as well as other experimental decisions (Section 4.3).

#### 4.1 Datasets

We used the test datasets made available by the WMT’19 News Translation shared task (Barrault et al., 2019). For each language pair, each source segment is associated with the following information:

- A reference translation in the target language (produced specifically for the task);
- The automatic translation outputted by each system competing in the task for that language pair;
- The average score obtained by each automatic translation, according to human assessments made by one or more human evaluators, in two formats: a raw score in [0;100] and a z-score in $[-\infty; +\infty]$. Not all the automatic translations received a human assessment;
- The number of human evaluators for each automatic translation (if there were any).

For brevity, we focused on five language pairs, listed in Table 1. The official top 3 systems for each pair, according to the average z-score, are shown in Table 2. Our choice of language pairs attempts to capture as many different phenomena as possible with the fewest pairs:

- **English → German** (en-de): This is the language pair with the most competitors and does not have a clear winning system (the winner differs depending on whether one considers the z-score or the raw score);
- **French → German** (fr-de): Unlike most language pairs, this pair features two languages other than English. Moreover, there is a strong imbalance between translations lacking human assessments and translations that received at least one assessment;
- **German → Czech** (de-cs): Besides featuring two languages other than English, this pair stands out as it was devised as an unsupervised task (i.e., English was used as a “hub” language);
- **Gujarati → English** (gu-en): This is one of the task’s low-resource language pairs (i.e., whose test set is half the size of most language-pairs in the task), and is one where there may

\(^2\)If multiple human assessments were made for the same translation, $score_{j,t}$ is the average of the scores received.
be more linguistic differences between the source and the target languages (e.g., different writing systems). Unlike en-de, there is a clear winner considering both raw and z-score. Moreover, three of the competing systems did not receive any human assessment on their translations;

- Lithuanian → English (lt-en): This is another low-resource language pair, with a rather competitive top 3. Unlike most language pairs, all the translations submitted by the competing systems for this pair received a human assessment.

For all these language pairs (except English → German), each segment was given an assessment score considering only the reference translation (and without access to the segment’s context within the document to which it belongs). For English → German, scores were given considering the source segment instead of the reference, and evaluators had access to the segment’s context within the document.

### 4.2 Human feedback

A key condition for applying online learning to this scenario is the availability of feedback. We use the human assessment raw scores in the test sets as a feedback source to compute the loss and update the weight of each MT system, as already suggested in Section 3. However, not all translations received human assessments (recall Table 1). To cope with this issue, we designed different variants of this loss function, following different fallback strategies:

- **human-zero**: If there is no human assessment for the current translation, a score of zero is returned (leading to an unchanged weight on that iteration);

### Table 1: Overview of the language pairs considered in our experiments.

<table>
<thead>
<tr>
<th></th>
<th>en-de</th>
<th>fr-de</th>
<th>de-cs</th>
<th>gu-en</th>
<th>lt-en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test set size (# segments)</td>
<td>1997</td>
<td>1701</td>
<td>1997</td>
<td>1016</td>
<td>1000</td>
</tr>
<tr>
<td>Competing systems</td>
<td>22</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>Human assessments coverage</td>
<td>86.80%</td>
<td>23.52%</td>
<td>62.94%</td>
<td>75.00%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

### Table 2: Top 3 performing systems for each language pair in the WMT’19 News Translation shared task (Barrault et al., 2019). The systems named “online-[letter]” correspond to publicly available translation services and were anonimized in the shared task.

<table>
<thead>
<tr>
<th></th>
<th>Top 3</th>
<th>z-score</th>
<th>Raw score</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-de</td>
<td>Facebook-FAIR (Ng et al., 2019)</td>
<td>0.347</td>
<td>90.3</td>
</tr>
<tr>
<td></td>
<td>Microsoft-sent-doc (Junczys-Dowmunt, 2019)</td>
<td>0.311</td>
<td>93.0</td>
</tr>
<tr>
<td></td>
<td>Microsoft-doc-level (Junczys-Dowmunt, 2019)</td>
<td>0.296</td>
<td>92.6</td>
</tr>
<tr>
<td>fr-de</td>
<td>MSRA-MADL (Xia et al., 2019)</td>
<td>0.267</td>
<td>82.4</td>
</tr>
<tr>
<td></td>
<td>eTranslation (Oravecz et al., 2019)</td>
<td>0.246</td>
<td>81.5</td>
</tr>
<tr>
<td></td>
<td>LIUM (Bougares et al., 2019)</td>
<td>0.082</td>
<td>78.5</td>
</tr>
<tr>
<td>de-cs</td>
<td>online-Y</td>
<td>0.426</td>
<td>63.9</td>
</tr>
<tr>
<td></td>
<td>online-B</td>
<td>0.386</td>
<td>62.7</td>
</tr>
<tr>
<td></td>
<td>NICT (Dabre et al., 2019)</td>
<td>0.367</td>
<td>61.4</td>
</tr>
<tr>
<td>gu-en</td>
<td>NEU (Li et al., 2019)</td>
<td>0.210</td>
<td>64.8</td>
</tr>
<tr>
<td></td>
<td>UEDIN (Bawden et al., 2019)</td>
<td>0.126</td>
<td>61.7</td>
</tr>
<tr>
<td></td>
<td>GTCOM-Primary (Bei et al., 2019)</td>
<td>0.100</td>
<td>59.4</td>
</tr>
<tr>
<td>lt-en</td>
<td>GTCOM-Primary (Bei et al., 2019)</td>
<td>0.234</td>
<td>77.4</td>
</tr>
<tr>
<td></td>
<td>tilde-nc-nmt (Pinnis et al., 2019)</td>
<td>0.216</td>
<td>77.5</td>
</tr>
<tr>
<td></td>
<td>NEU (Li et al., 2019)</td>
<td>0.213</td>
<td>77.0</td>
</tr>
</tbody>
</table>

3Although we assume an absolute scale of scores in [0;100] in our experiments, our approach could be applied to any other level of granularity.
• human-avg: If there is no human assessment for the current translation, the average of the previous scores received by the system behind that translation is returned as the current score;

• human-comet: If there is no human assessment for the current translation, the COMET score (Rei et al., 2020a) between the translation and the pair source/reference available in the corpus is returned as the current score.

We pre-trained this automatic metric on the datasets of previous shared tasks (WMT’17 (Bojar et al., 2017) and WMT’18 (Bojar et al., 2018)). Thus, for most translations, it displays a small difference regarding the existing human scores (see Fig. 2 for the case of en–de). Moreover, this metric correlates better with ratings by professional translators than the WMT scores (Freitag et al., 2021).

4 We trained this metric from scratch following the hyperparameters described in Rei et al. (2020b), except that we used the raw scores instead of the z-normalized scores.

5 Results and discussion

In order to observe whether (and how soon) our online approach converges to the best systems, we report the overlap between the top $n = 1, 3$ systems with greatest weights according to our approaches, $s_n$, and the top $n = 1, 3$ systems according to the shared task’s official ranking, $s_n^\ast$, at specific iterations:

$$\text{top}_n = \frac{|s_n \cap s_n^\ast|}{n}, \, n = 1, 3$$

We preferred this metric over a rank correlation metric, as we are focused on whether our online approach follows the performance of the best MT systems. In a realistic scenario (e.g., a Web MT service), a user would most likely rely solely on the main translation returned, or would at most consider one or two alternative translations. Moreover, due to the lack of a large enough coverage of human assessments, the scores obtained in the shared task are not reliable enough to discriminate between similarly performing systems.

Starting with en–de (Table 3), this was the language pair for which our approach appears to be the least successful, since, for most of the iterations examined, it failed to converge to the best system. Even so, it managed to converge to the top 3 systems, doing so particularly early in the learning process (50 iterations) when using EWAF with human-avg and human-comet as loss functions (i.e., when using fallback scores). Recall that, for this language pair, there were different official winning systems depending on whether one considers the z-score or the raw score (recall Table 2); since we use the raw score as the loss function, it is expectable that our approach does not necessarily converge to the winner according to the z-score.

For fr–de (Table 4), our online approach often converges to the top 3 systems (or a subset of them) throughout the learning process (even at just 10 iterations), and it also converges to the best system when using EWAF with human-comet. This is a particularly interesting result if we recall that, out of the five pairs considered, fr–de had the lowest coverage of human assessments by far (see Table 1), thus suggesting that using COMET may be an adequate fallback strategy.

For de–cs (Table 5), we can see that, regardless of the algorithm and loss function used, there is an overlap of at least one system between our top 3 and the shared task’s official top 3, after going through
only as few as 10 iterations (despite a considerable lack of human assessments in this language pair). We can also see that the human-comet loss function is the most successful overall, which reinforces the idea that COMET may be an appropriate fallback metric in the absence of human scores for a given translation. Since this is the language pair for which there seems to be a more similar performance across different algorithms and loss functions, we also report the weight evolution plots for this pair in order to inspect what changes depending on the algorithm and fallback strategy used. Looking at EWAF combined with the human-zero loss function (Fig. 3), one can see a rather irregular evolution for the weights of the top systems, which may be explained by the distribution of the translations lacking human assessments across different systems and learning iterations. Using the human-avg loss function (Fig. 4) allows for a more monotonous evolution, by rewarding the systems.

The plots for the remaining pairs can be found in App. A.
Table 6: Overlap ratios of top 1 and top 3 systems in common between the online approaches and the official ranking for gu-en. Recall that there were three systems competing on this language pair that did not receive human assessments at all (thus, using human-avg yields the same results as using human-zero).

<table>
<thead>
<tr>
<th>Iteration</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>500</th>
<th>1000</th>
<th>1016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top</td>
<td>1.00</td>
<td>0.67</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>EWAF</td>
<td>1.00</td>
<td>0.67</td>
<td>0.00</td>
<td>0.67</td>
<td>0.00</td>
<td>0.67</td>
</tr>
<tr>
<td>human-zero</td>
<td>0.00</td>
<td>0.33</td>
<td>0.00</td>
<td>0.67</td>
<td>0.00</td>
<td>0.33</td>
</tr>
<tr>
<td>human-comet</td>
<td>0.00</td>
<td>0.33</td>
<td>0.33</td>
<td>1.00</td>
<td>0.33</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 7: Overlap ratios of top 1 and top 3 systems in common between the online approaches and the official ranking for lt-en. Recall that this was the only language pair for which all the translations received at least one human assessment, thus there is no need to use a fallback loss function.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top</td>
<td>1.00</td>
<td>0.67</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>EWAF</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.67</td>
<td>0.00</td>
</tr>
<tr>
<td>human-zero</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.33</td>
</tr>
<tr>
<td>human-comet</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

As for gu-en (Table 6), our approach (using EWAF with human-zero) converges to the best system and to a subset of the top 3 within just 10 iterations; on the other hand, using human-comet does not do as well as not using a fallback strategy, at least when combined with EWAF. However, recall that, for this pair, there were systems that did not receive any human assessments at all for their translations (that being the reason why we do not report human-avg for this pair: the resulting weights end up being the same as when using human-zero). One of the systems that did not receive any human assessments, online-B, ended up receiving high COMET scores, thus leading to a weaker overlap between the online approach ranking and the official ranking.

Finally, for lt-en (Table 7) we only report the human-zero loss function, since this is the only pair for which there are human assessments for all translations. Interestingly, the online approaches do not do well as quickly as for other pairs, but eventually get there (within 100 to 500 iterations).

To sum up these results: although factors like the coverage of human assessments or the combinations of online algorithm and loss function used influence how well our approach does, we can still conclude that using an online learning approach allows to converge to the top 3 systems according to the official ranking (or at least to a subset of them) in just a few hundred iterations (and, in some cases, in just a few dozens of iterations) for all the language pairs considered.

6 Related work

6.1 WMT’19 News Translation Shared Task

Every year, since 2006, the Conference on Machine Translation (WMT) is responsible for organizing several shared tasks where participants push the limits of MT and MT evaluation (Koehn and Monz, 2006; Barrault et al., 2020). In the News Translation shared task, participants submit the out-
puts of their systems that are then evaluated by a community of human evaluators using Direct Assessment scores (Graham et al., 2013). Thus, the winner is the system that achieves the highest average score. For WMT’19 (Barrault et al., 2019), most of the competing systems followed a Transformer architecture (Vaswani et al., 2017), with the main differences among them being: (i) whether they considered document-level or only sentence-level information; (ii) whether they were trained only on the training data provided by the shared task, or on additional sources as well; (iii) whether they consisted of a single model or an ensemble.
6.2 Online learning for Machine Translation

There has been a number of online learning approaches applied to MT in the past, mainly in Interactive MT and/or post-editing MT systems. However, most approaches aim at learning the parameters or feature weights of an MT model (Mathur et al., 2013; Denkowski et al., 2014; Ortiz-Martínez, 2016; Sokolov et al., 2016; Nguyen et al., 2017; Lam et al., 2018) or fine-tuning a pre-trained model for domain adaptation (Turchi et al., 2017; Karimova et al., 2018; Peris and Casacuberta, 2019). Even in cases where the MT model is composed of several sub-models (e.g., Ortiz-Martínez (2016)), the goal is to online learn each sub-model’s specific parameters (while our learning goal is the weights of each system in an ensemble). Another key difference between these approaches and ours is that most of them use human post-edited translations as a source of feedback. The exceptions to this are the systems competing for WMT’17 shared task on online bandit learning for MT (Sokolov et al., 2017), as well as Lam et al. (2018), who use (simulated) quality judgments.

The most similar proposal to ours is that of Naradowsky et al. (2020), who ensemble different MT systems and dynamically select the best one for a given MT task or domain using stochastic multi-armed bandits and contextual bandits. The bandit algorithms learn from feedback simulated using a sentence-level BLEU score between the selected automatic translation and a reference translation. Thus, to the best of our knowledge, we are the first to frame the MT problem as a problem of prediction with expert advice and adversarial multi-armed bandits in order to combine different systems into an ensemble that converges to the performance of the best individual systems, simulating the human-in-the-loop by using actual human assessments (when available).

7 Conclusions and future work

We proposed an online learning approach to address the issue of finding the best MT systems among an ensemble, while making the most of existing human feedback. In our experiments on WMT’19 News Translation datasets, our approach converged to the top-3 systems (or a subset of them) according to the official shared task’s ranking in just a few hundred iterations for all the language pairs considered (and just a few dozens in some cases), despite the lack of human assessments for many translations. This is a promising result, not only for the purpose of reducing the human evaluations required to find the best systems in a shared task, but also for any MT application that has access to an ensemble of multiple independent systems and to a source of feedback from which it can learn iteratively (e.g., Web translation services).

Yet, our approach is limited by the quality of the collected human judgments. For future work, we plan to combine online learning with a more reliable human metric, such as the Multidimensional Quality Metric (MQM) (Lommel et al., 2014), so that we can focus on the quality of the assessments instead of their quantity.

Acknowledgments

This work was supported by: Fundação para a Ciência e a Tecnologia (FCT) under references UIDB/50021/2020 (INESC-ID multi-annual funding) and UIDB/00214/2020 (CLUL), as well as under the HOTSPOT project with reference PTDC/CCI-COM/7203/2020; Air Force Office of Scientific Research under award number FA9550-19-1-0020; P2020 program, supervised by Agência Nacional de Inovação (ANI), under the project CMU-PT Ref. 045909 (MAIA), Vânia Mendonça was funded by an FCT grant with reference SFRH/BD/121443/2016.

The authors would like to thank the reviewers for their valuable comments, and to Soraia M. Alarcão for kindly proof-reading this document.

References


Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn,


A Weight evolution (all language pairs)

Here we present the weight evolution per MT system for all the combinations of language pairs, learning algorithms (EWAF or EXP3), and loss functions (human-zero, human-avg, or human-comet, when applicable) – except for those combinations that are already part of the main document.

A.1 English → German (en-de)

Figure 7: EWAF with human-zero loss. Recall that, for this language pair, the official top 3 systems were Facebook-FAIR, Microsoft-sent-doc, and Microsoft-doc-level.

Figure 8: EWAF with human-avg loss.

Figure 9: EWAF with human-comet loss.

Figure 10: EXP3 with human-zero loss.

A.2 French → German (fr-de)

Figure 11: EXP3 with human-avg loss.

Figure 12: EXP3 with human-comet loss.

Figure 13: EWAF with human-zero loss. Recall that, for this language pair, the official top 3 systems were MSRA-MADL, eTranslation, and LIUM.

Figure 14: EWAF with human-avg loss.

Figure 15: EWAF with human-comet loss.

Figure 16: EXP3 with human-zero loss.
A.3 German → Czech (de-cs)

Figure 19: EXP3 with human-avg loss. Recall that, for this language pair, the official top 3 systems were online-Y, online-B, and NICT.

A.4 Gujarati → English (gu-en)

Figure 21: EWAF with human-zero loss. Recall that, for this language pair, the official top 3 systems were NEU, UEDIN, and GTCOM-Primary.

A.5 Lithuanian → English (lt-en)

Figure 25: EWAF with human-zero loss. Recall that, for this language pair, the official top 3 systems were GTCOM-Primary, tilde-nc-nmt, and NEU.