Bias Mitigation in Machine Translation Quality Estimation

Hanna Behnke
Imperial College London
hanna.behnke20@ic.ac.uk

Marina Fomicheva
University of Sheffield
m.fomicheva@sheffield.ac.uk

Lucia Specia
Imperial College London
l.specia@ic.ac.uk

Abstract

Machine Translation Quality Estimation (QE) aims to build predictive models to assess the quality of machine-generated translations in the absence of reference translations. While state-of-the-art QE models have been shown to achieve good results, they over-rely on features that do not have a causal impact on the quality of a translation. In particular, there appears to be a partial input bias, i.e., a tendency to assign high-quality scores to translations that are fluent and grammatically correct, even though they do not preserve the meaning of the source. We analyse the partial input bias in further detail and evaluate four approaches to use auxiliary tasks for bias mitigation. Two approaches use additional data to inform and support the main task, while the other two are adversarial, actively discouraging the model from learning the bias. We compare the methods with respect to their ability to reduce the partial input bias while maintaining the overall performance. We find that training a multitask architecture with an auxiliary binary classification task that utilises additional augmented data best achieves the desired effects and generalises well to different languages and quality metrics.

1 Introduction

Despite the great advances of Machine Translation (MT) models over the past years, the adequacy and fluency of the translations cannot be guaranteed. Without access to a gold-standard reference translation, it can be difficult to validate the reliability of the MT model's predictions. To address this issue, the field of MT Quality Estimation (QE) emerged, aiming to develop models that can approximate the quality of machine-generated translations in a scalable way. However, recent research suggests that state-of-the-art QE approaches tend to over-rely on features that do not have a causal impact on the quality of a translation. In particular, there appears to be a partial input bias, i.e. a tendency to assign high quality scores to translations that are fluent and grammatical, even though they do not resemble the actual meaning of the source (Sun et al., 2020).

Building upon these findings, the objective of our work is to characterise and, most importantly, mitigate the partial input bias of QE models. We focus on the use of auxiliary training tasks to specifically target the observed biases while avoiding strong modifications of the original model as well as the expensive collection and manual labelling of additional training data. Our efforts concentrate on testing and improving *MonoTransQuest*, the best-performing architecture in the shared task on sentence-level QE hosted as part of the *Fifth Conference on Machine Translation (WMT 2020)* (Specia et al., 2020). We work with the recently published multilingual QE dataset *MLQE-PE* (Fomicheva et al., 2020), allowing us to test the generalisability of our approaches across different languages and quality scores.

Our main **contributions** are as follows:

- **Bias analysis.** We expand on previous research which suggested the partial input bias in QE and find that the model as well as the annotators tend to over-rate the quality of fluent but inadequate translations.
- **Bias mitigation.** To the best of our knowledge, we are the first to explore the mitigation of biases with auxiliary tasks in the field of QE. We group our approaches into four categories: Multitask training with mixed languages, multitask training with additional augmented data, training with adversarial tasks and training with debiased focal loss.
- **New architectures.** We implement and compare several multitask architectures and find that iteratively training the tasks with two optimisers is better suited for our objective than backpropagating a weighted sum of the losses. Further, we reformulate focal loss for regres-

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1475 - 1487
May 22-27, 2022 ©2022 Association for Computational Linguistics

sion tasks, a technique that is traditionally based upon the cross-entropy loss.

- **Results.** Utilising the multitask architecture, we successfully reduce the partial input bias while maintaining the same performance as the benchmark model and examine the best model's robustness.

In the subsequent sections, we first present related work, followed by the analysis of the partial input bias. Building upon the findings, we explain the four bias mitigation approaches in Section 4 and discuss the results in Section 5.

2 Related Work

2.1 Machine Translation Quality Estimation

QE is an area of research concerned with the development of models for the prediction of the quality of machine-generated translations when gold standard translations are not available. QE is normally addressed as a supervised machine learning task, which may take as input general information from the source and translated texts, as well as from the MT system. The quality is typically assessed at sentence level, but word- and document-level QE are also possible (Specia et al., 2018, pp. 2). Sentence-level QE has evolved from the first feature-heavy prediction models (Blatz et al., 2004) to neural architectures such as RNNs and Transformers (Vaswani et al., 2017), which accelerated the developments in the field by reducing the work of manual feature engineering and improving contextual representations (Kim et al., 2017; Wang et al., 2018; Fan et al., 2019).

A prominent state-of-the-art QE architecture is MonoTransQuest, proposed by Ranasinghe et al. (2020). It builds upon XLM-R, a popular pre-trained cross-lingual language model with a good ability to generalise to low-resource languages (Conneau et al., 2020). MonoTransQuest achieved the best results for sentence-level direct assessment score prediction in the WMT 2020 shared task on QE (Specia et al., 2020).

Sun et al. (2020) showed that QE models like MonoTransQuest have a tendency to over-rely on spurious correlations, which is partially due to skewed label distributions and statistical artifacts in QE datasets. In particular, they show the existence of a partial input bias, i.e. the tendency to predict the quality of a translation based on just the target sentence (Poliak et al., 2018). While the fluency and grammatical correctness of the output is a factor influencing the quality, the original meaning should be preserved, which is only possible if the model takes both source and target into consideration. Following their work, in an attempt to reduce statistical artifacts, MLQE-PE (Fomicheva et al., 2020) – a new QE dataset diversifying the topics and languages covered – was created, which forms the basis of this work and will be described in more detail in Section 3.1.

2.2 Bias Mitigation with Auxiliary Tasks

We define auxiliary tasks in a broad sense, using the term to refer to settings where a main task is trained alongside one or more helper tasks used to improve the main task's performance and generalisability. Most commonly, the tasks are trained in a multitask-setting, where some layers are shared across the tasks and some layers are task-specific. The auxiliary tasks can either be *related* to the main task or *adversarial* (Ruder, 2017). In addition, we consider the concept of *debiased focal loss*, where the main and auxiliary task are trained in separate models which are connected via the loss function (Karimi Mahabadi et al., 2020).

Related Tasks: In settings where the data is limited, noisy or high-dimensional, using additional tasks is a way of introducing an inductive bias that prevents the model from overfitting to noise (Caruana, 1997). In addition, the model might be able to use new features that were learned through an auxiliary task for the main task as well (Ruder, 2019). MT models, for example, have been shown to benefit from auxiliary tasks such as named entity recognition, part-of-speech tagging and dependency parsing (Niehues and Cho, 2017; Kiperwasser and Ballesteros, 2018).

Adversarial Tasks: Adversarial tasks can be used to actively discourage the model from overfitting to domain-specific, spurious cues. The technique was introduced by Ganin and Lempitsky (2015) and used for domain adaptation. More recently, it has been successfully used to reduce partial input biases in different fields of NLP, such as natural language inference (NLI) (Belinkov et al., 2019; Stacey et al., 2020) and visual question answering (VQA) (Ramakrishnan et al., 2018). The core idea is to train the auxiliary task using just the partial input. During backpropagation, the gradient is reversed. Consequently, the shared layers are updated such that the adversary's loss is maximised;

the undesired behaviour is penalised. The methodology chapter illustrates the architectural design in more detail.

Debiased Focal Loss: Another approach that has recently been used to mitigate known biases, particularly partial input biases, is debiased focal loss. The notion of focal loss was first introduced by Lin et al. (2017) as a means to improve classification results on imbalanced classes by weighing down the impact of samples that the model had already learned to classify well. In the field of NLI, Karimi Mahabadi et al. (2020) have shown that it is possible to adapt the notion of focal loss to mitigate partial input biases. They train the main model alongside a bias model that learns to predict the label based on the hypothesis only. In this scenario, the bias model's predictions are used to weight the main model's cross-entropy loss. Intuitively, samples that are classified well by the bias model are weighted down so that the main model primarily learns from less biased inputs. The bias model is updated separately and discarded after training.

3 Bias Analysis

In the following, we will describe the dataset and baseline model used, show benchmark results and analyse the partial input bias in more detail.

3.1 Dataset

We work with the MLQE-PE dataset (Fomicheva et al., 2020) which was specifically designed for the training of MT QE models. Published in 2020, it formed the basis for the WMT 2020 and 2021 shared tasks on Quality Estimation (Specia et al., 2020).[1] It consists of 6 high-, mid- and low-resource language pairs which originate from Wikipedia articles: English-German and English-Chinese, Romanian-English and Estonian-English as well as Nepalese-English and Sinhala-English. A seventh dataset, Russian-English, was curated from Reddit posts and WikiQuotes. The translations were generated using Transformer-based Neural MT models. For each language, 9000 sentence pairs (7000 train, 1000 dev, 1000 test) were annotated on two different scales:

- **Human-targeted Translation Edit Rate (HTER)**: Each sentence-pair was edited by two independent translators. The reported HTER score is the averaged edit rate comparing the machine-generated translations and the post-edited versions. The score ranges between 0 (perfect translation) to 1 (everything was edited).
- **Direct Assessment Scores (DA)**: Each sentence pair was judged on a scale from 0-100 by at least 3 evaluators. The reported DA score is the mean of the individual judgements. Different than the HTER scores, the DA scale provides a measure of the severity of the errors, where inadequate (i.e. non-meaning preserving) translations should not receive a score higher than 70, even if only one word is incorrect.

3.2 Benchmark

We use the XLM-R based architecture MonoTransQuest as our baseline model, which fine-tunes XLM-R for sentence-level QE (Ranasinghe et al., 2020). While there are alternative candidates with a good performance on QE tasks, MonoTransQuest was chosen for several reasons: State-of-the-art performance, availability and replicability (all hyperparameters and the source code are open-sourced), as well as the generic design of the architecture which is transferable to related NLP domains.

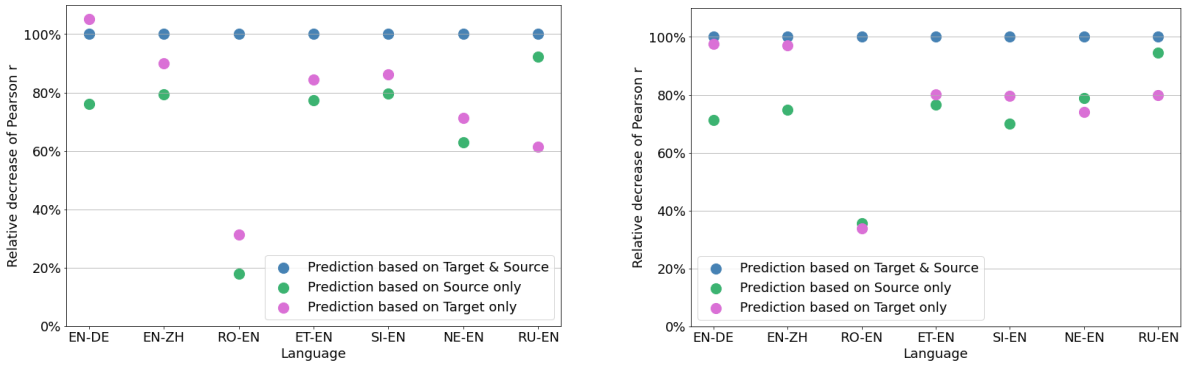
We train separate MonoTransQuest models for each combination of language pair and quality score using the originally proposed architecture and fine-tuned hyperparameters specified in the TransQuest GitHub repository.[2] All experiments were conducted on a 16GB Nvidia Tesla P100 GPU and averaged across five trainings on the seeds 555, 666, 777, 888 and 999. Our results are shown in Table 3 in the Appendix. In QE, the best practice is to use Pearson's r to measure performance (Specia et al., 2018, pp. 58). Most notably, the Pearson correlation between the predictions and the labels is lowest for the high-resource languages English-German and English-Chinese. This has also been observed in the QE shared task findings (Specia et al., 2020). A possible explanation is the high average quality of the generated translations, making the labelling significantly harder and the annotations less consistent, i.e. more noisy.

3.3 Partial Input Bias

We examine the partial input bias by training the model on the combined representation of source

[1] The train, dev and *test20* test sets are available via https://github.com/sheffieldnlp/mlqe-pe

[2] https://github.com/TharinduDR/TransQuest



(a) DA predictions

(b) HTER predictions

Figure 1: Relative decrease of the correlation between prediction and labels when testing with source or target, only.

and target and testing how the performance changes when the prediction is based on only one of the two. If the performance does not significantly decrease, the model has likely learned to base its predictions mostly on one part of the input. Figure 1 shows the results from this experiment. A clear target sentence bias can be observed for the English-German and English-Chinese language pairs. One reason could be the good quality of the translations that MT systems generate for high-resource languages: The occurrence of adequacy errors is lower, so that the target sentence may suffice for a decent prediction. In contrast, the mid-resource Romanian-English model, which shows the best overall performance, appears to be most dependent on both inputs. Figure 1 shows a clear performance deterioration when the model is tested on just the source or target sentence. One particularity of the RO-EN dataset is the high abundance of fluent, but clearly inadequate translations and hallucinations which require both the source and translation to be detected (Specia et al., 2020). The Russian-English dataset is an exception where the source sentence is a good predictor for the translation quality, most likely due to the distinct nature of Reddit data and WikiQuotes (both user-generated). This source sentence bias could best be mitigated by curating a new dataset which is why we chose not to focus our efforts on the Russian-English dataset.

To further examine the nature of the partial input bias, an in-depth analysis of the strongly affected English-German translations was conducted. In particular, the aim was to better understand how MonoTransQuest, but also the annotators, judge the quality of fluent but inadequate translations. To achieve this, one of the authors, a German native speaker, manually annotated translations in the test set that are grammatically correct but do not preserve the meaning of the source.[3]

In total, 145 out of 1000 translations were marked as fluent but inadequate. A key takeaway from the labelling process was that it is not only the models that have a partial input bias – human annotators clearly seem to over-rely on the target fluency, too. Even if the instructions clearly specify that a DA score below 70 should be assigned to inadequate translations,[4] annotators tended to give higher scores if the sentence was fluent and appeared logical. Figure 2 shows that more than half of the fluent but inadequate translations were given a score higher than 70, with an average rating of 81.[5] A likely reason is that adequacy-related mistakes are easy to miss when considering several quality factors, i.e. spelling, grammar and content, at the same time.

4 Methodology

Based on the bias analysis, our goal is to find an effective and feasible way to reduce the impact of spurious correlations and overly dominant features. As outlined in the previous section, the two high-resource datasets (EN-DE and EN-ZH) clearly show the strongest partial input bias. They will therefore be at the centre of the bias mitigation efforts. All four methods presented hereinafter share the core idea of using auxiliary tasks to achieve this aim: The main task – QE – is combined with helper tasks designed to reduce known

[3]The annotated dataset is available via https://github.com/agesb/TransQuest

[4]The DA annotation guidelines used in the MLQE-PE data dictate that a score in 70–90 indicates a translation that closely preserves the semantics of the source sentence.

[5]The HTER score was not examined in this analysis since it does not explicitly account for the adequacy of the translation.

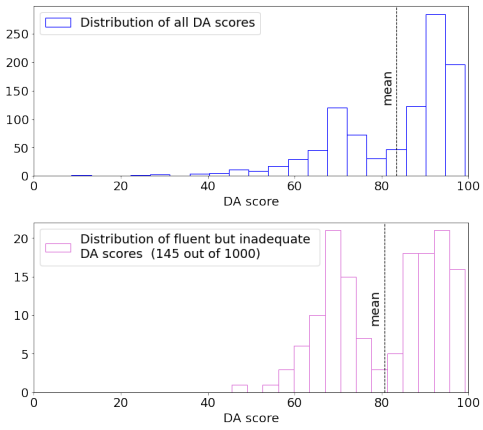


Figure 2: DA label and prediction distribution of fluent but inadequate translations.

biases. At test time, the auxiliary tasks can be discarded. Hereinafter, we introduce four approaches and the corresponding model architectures. The first two methods are tailored to combat the biased behaviour by supporting the model with additional data. In contrast, the two alternative, restrictive approaches actively penalise the model for learning unwanted behaviour. We define three criteria to ensure comparability between the approaches: A good solution should 1) mitigate the observed biases, 2) retain the prediction quality of the benchmark model, and 3) avoid computational overhead and interference with the original model's design.

4.1 Supportive Approaches

We experiment with two different supporting tasks, each combining the main task and the auxiliary task in a multitask setup. The first approach is to train with different language pairs, aiming to transfer information between the language domains. Instead of mixing the languages arbitrarily, we build upon the bias analysis and examine if using a less biased language (RO-EN) to train the auxiliary task can help to reduce biases in the main task (EN-DE or EN-ZH). The bias analysis clearly showed that the models trained on the RO-EN dataset performed poorly when using just the source or target as input, indicating that the predictive power of the individual sentences is low. Thus, the incentive for the multitask model to over-rely on the target should be reduced. In this scenario, both tasks are regression problems and optimise the MSE loss.

The second approach is to collect additional translations originating from the same topic and language domain and use it as the input for the auxiliary task. We choose WikiMatrix (Schwenk et al., 2021), a large parallel sentence corpus based on Wikipedia articles, as data source for the experiments. Without further preprocessing, the vast majority of these sentence pairs would qualify as good translations. While labelling on a continuous scale would require manual annotations, augmenting the data to achieve "bad" translations is more feasible. Hence, we augment 50% of the data to obtain bad translations. We experiment with two augmentation strategies: First, we shuffle the sentences to create mismatched sentence pairs. Second, we augment the sentence to mimic fluent but inadequate translations as seen in the original MLQE-PE dataset and discussed in Section 3.3. To do so, we implement a contextual augmentation pipeline that uses a language model (XLM-R) to replace 30% of the nouns, adjectives, verbs and adverbs such that the meaning of the sentence is changed while the grammatical correctness is preserved in the majority of cases.[6] In both cases, the main task optimises the MSE loss, and the auxiliary task is a binary classification problem using the binary cross-entropy loss.

4.2 Restrictive Approaches

We experiment with two setups that directly penalise the biased behaviour. First, we combine the main task with an adversarial task in a multitask architecture. Intuitively, the adversary is incentivised to predict the quality scores based on the target sentence only. The shared layers, on the other hand, are penalised for learning a mapping between target sentence and scores. The risk of working with an adversarial task setup is that it optimises towards eliminating all cues associated with the adversary. In QE, however, the target sentence provides relevant information, such as grammar and spelling. As a result, the overall model performance might suffer. As an alternative to training with adversarial tasks and a multitask architecture in general, we repurpose the concept of debiased focal loss for regression. While model architecture and training method are different, the underlying idea to use the partial input based predictions to influence the learning remains the same. The subsequent section explains the multitask architecture used for the first three approaches as well as the re-formulated debiased focal loss technique in more detail.

[6]The augmentation pipeline was published as part of the NL-Augmenter library (Dhole et al., 2021): https://github.com/GEM-benchmark/NL-Augmenter/tree/main/transformations/contextual_meaning_perturbation

4.3 Architecture & Training

4.3.1 MultiTransQuest

To realise the first three approaches, we propose the architecture *MultiTransQuest*, expanding on the MonoTransQuest baseline. The pre-trained language model XLM-R remains at the core and is entirely shared between tasks. The two key changes affect the final layers and the optimisation strategy: Firstly, we exchange the original prediction head to support multiple tasks. As illustrated in Figure 3, the final layers and loss functions are separate per task, thus allowing the mixing of regression and classification tasks. The figure exemplarily shows the adversarial setup, where the gradients are reversed during back-propagation, i.e. weighted with -1. For the two supportive tasks, we use the same setup but remove the weighted gradient layers and adjust the input and loss function for the auxiliary tasks accordingly. We experiment with different numbers of shared and separate layers. Secondly, we adapt the training procedure to support multiple tasks. The data loader is designed so that it alternates between the tasks per training step, with each batch containing only samples for one task which are then passed through the shared layers and the corresponding task-specific layers. We compare two optimisation strategies:

- Training the tasks in turns, where backpropagation is performed per batch and task. Each task works with a separate AdamW optimizer to avoid averaging gradients across tasks.
- Performing one forward pass for every task and combining the calculated losses as a weighted sum which is backpropagated through all layers using a single optimizer.

4.3.2 Debiased Focal Loss Architecture

In contrast to the previously discussed multitask approaches, debiased focal loss enables a complete separation of the main model and bias model, thus requiring no changes to the core MonoTransQuest architecture. To the best of our knowledge, (debiased) focal loss has only been applied to classification tasks so far as it explicitly modifies the cross-entropy loss function. Since our QE task is formulated as a regression problem, we attempt to find an equivalent strategy to weigh down biased examples when working with MSE loss. In our scenario, the bias model is trained on partial inputs, receiving the translated sentence only. The better

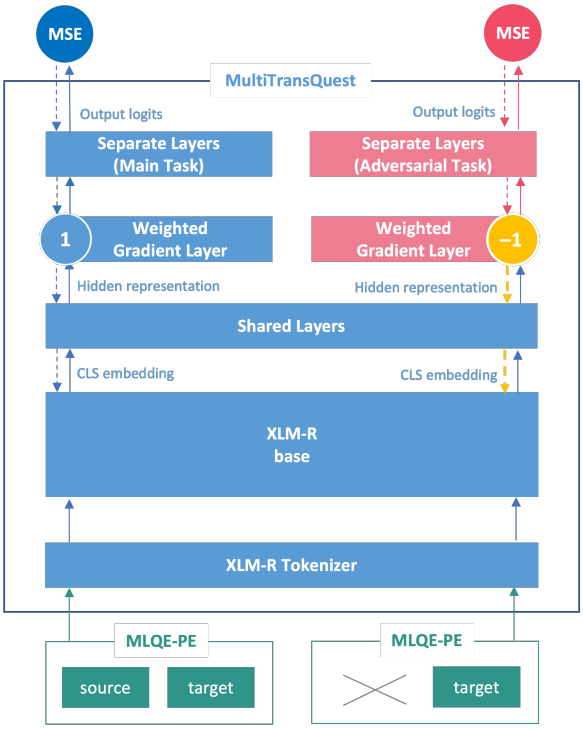


Figure 3: Multitask architecture with gradient reversal.

the bias model's prediction, the lower the MSE and the more biased the sample. In line with the original debiased focal loss idea, we can therefore use the bias model's loss as an indication for the bias per sample.

As the MSE loss can vary greatly during training, we decide against training both models in an end-to-end approach. First, the trained bias model is used to predict the respective quality scores for the training set, using only the target. Next, the absolute error for each of the training samples is calculated. We use the error to approximate the partial input bias: The lower the error, the easier it is for the bias model to predict the sample's quality score correctly. To control the scale of the weights, we normalise the error value between 0 and 1. The resulting weights w are used to scale the MSE loss of the main model f_M before backpropagation. We use the hyperparameter β to exponentially scale the loss (Eq. 1). We further experiment with a sigmoid-shaped function scaled between 0 and 1 (Eq. 2).

$$DFL = w^{\beta} \left(f_M^{y_i} \left(\boldsymbol{x_i} \right) - \hat{y}_i \right)^2 \quad (1)$$

$$DFL = \frac{1}{1 + \left(\frac{w}{1-w} \right)^{-\beta}} \left(f_M^{y_i} \left(\boldsymbol{x_i} \right) - \hat{y}_i \right)^2 \quad (2)$$

Data	Experiment	DA			HTER		
		r	MSE	***r target***	***r***	MSE	***r target***
EN-DE	***benchmark***	0.3695±0.03	0.0239	0.4189	0.4734±0.01	0.0308	0.4555
	bilingual	0.3748±0.05	0.0285	0.2307	0.4718±0.01	0.0334	0.4103
	augmented	0.4163±0.04	0.0299	-0.0822	0.4512±0.01	0.0359	0.2279
	adversarial	0.2086±0.08	0.0215	-0.0926	0.4429±0.01	0.0334	0.1312
	focal	0.3184±0.05	0.0189	0.3148	0.4470±0.02	0.0312	0.4152
EN-ZH	***benchmark***	0.4249±0.01	0.0246	0.3746	0.3337±0.01	0.0792	0.3103
	bilingual	0.4008±0.03	0.0317	0.3282	0.3222±0.01	0.0833	0.2623
	augmented	0.3998±0.02	0.0300	0.1283	0.3328±0.02	0.0911	0.2237
	adversarial	0.3899±0.01	0.0289	0.0474	0.2824±0.01	0.0868	0.0695
	focal	0.4255±0.01	0.0437	0.3988	0.3322±0.01	0.0748	0.2969

Table 1: **Results**. Comparison of the four bias mitigation approaches. Column *r* shows the mean Pearson correlation of labels and predictions and the standard deviation over 5 runs, each training for 3 epochs = 15 minutes. Column *MSE* is the average mean squared error. Column *r target* measures the performance when testing on the target sentence only and thus approximates the bias mitigation effect, where a smaller correlation is better.

5 Results

In the following, we present and discuss the results of the experiments conducted. Based on the analysis in Section 3.3, the experiments concentrate on the two most biased datasets English-German and English-Chinese, each in combination with the DA and HTER scores. For each of the four sections, we assess different hyperparameter configurations on the EN-DE validation set. A configuration is considered to be good if the bias is reduced and the overall performance is at least maintained. The most promising variant is then evaluated on the EN-DE and EN-ZH test set, to see if the method generalises across language domains. Finally, we compare the four methods against one another and provide further analyses on the robustness of the best-performing model. [7]

5.1 Hyperparameters and Design Choices

Within each of the four approaches, we experiment with different hyperparameter configurations and design choices. While each setup requires individual fine-tuning, observed trends, backed by Table 4, 5, 6, 7 and 8 in the Appendix, include:

- For the multitask architecture, training the tasks in turns with separate optimisers results in a good balance between bias reduction and maintaining performance. Backpropagating the weighted loss is also possible, but requires more task-specific fine-tuning.
- For supportive auxiliary tasks, more separate layers, i.e. a larger degree of freedom, and a larger batch size improve the performance, for adversarial tasks the opposite is the case.
- When augmenting additional WikiMatrix data, shuffling the sentence pairs achieves better results than mimicking fluent but inadequate translations with contextual augmentation.
- The effect of the debiased focal loss technique is limited. A sigmoid-shaped weight distribution does not improve the results.

5.2 Comparison of the Four Approaches

Table 1 summarises the results obtained for each of the four methods. With respect to the choice of architecture, MultiTransQuest, used for methods 1-3, reduces the partial input bias more effectively than MonoTransQuest trained with focal loss. A key advantage of the multitask architecture is that the model is able to learn a balance between the tasks. In contrast, the degree of freedom is significantly limited for the focal loss architecture, where the main hyperparameter is how to scale the weights. We believe that this limitation is what makes the model even more sensitive to the inseparability of the bias and helpful features.

Contrasting the multitask-training with related or adversarial tasks, we find that the two supportive methods maintain a solid performance across all four constellations, while also reducing the bias. Compared to this, the adversarial approach gen-

[7] For reproducibility of the experiments, the source code incl. configurations is published under https://github.com/agesb/TransQuest. All hyperparameters not explicitly mentioned in the paper were kept constant.

eralises less well, despite its successful application in NLI and VQA. We hypothesise that this discrepancy is rooted in the nature of the partial inputs: In VQA as well as NLI, the task can only be solved when considering both question and image or premise and hypothesis, respectively. In contrast, the translation provides information that is valuable for the QE model regardless of the source, such as the fluency of the generated sentence. Hence, it is difficult to isolate the bias from valuable information, an assumption that both adversarial training and the focal loss technique rely on. Without an unbiased reference dataset (which is hard to acquire due to the subjective nature of the annotation process) the line between desired information and bias is difficult to quantify. The lower the correlation between the existence of the bias and the performance of the adversarial task, the noisier the feedback that is propagated into the shared layers.

The best trade-off between overall performance and bias reduction is achieved with MultiTransQuest when combining the main task with a binary classification task trained on shuffled WikiMatrix data. The binary classification task is simple to learn, yet impossible to solve without paying equal attention to source and translation. For better illustration of the model behaviour and improvements, Figure 6 in the Appendix directly compares the performance and bias reduction achieved by the best model to the benchmark. In addition, Figures 7 and 8 show the distribution of DA and HTER predictions generated by the debiased model.

Since the reduction of the performance on the target sentence is only considering the reduction of the partial input bias, we additionally aim to test the model's ability to generalise better on datasets that barely exhibit the partial input bias. As a feasible alternative to collecting an unbiased reference dataset in the same language domain, we assess the models' robustness in a zero-shot setting on less biased RO-EN data. As elaborated on in Section 3.3, the RO-EN dataset provokes the partial input bias significantly less than the other language pairs. Consequently, a model with reduced partial input bias should perform better when tested on the dataset, indicating improved robustness. We train the MonoTransQuest benchmark and debiased MultiTransQuest architecture on the EN-DE and EN-ZH datasets and use these models to predict the respective scores on the RO-EN dataset. Since this is an out-of-domain setting, we do not expect the models to reach a performance that can compete with the models trained on Romanian-English data. However, the debiased MultiTransQuest models should outperform MonoTransQuest in this zero-shot scenario, which is indeed the case as can be seen from Table 2.

	EN-DE model		EN-ZH model	
	DA	**HTER**	**DA**	**HTER**
MonoTQ	0.3756	0.3466	0.494	0.3650
MultiTQ	0.5601	0.3543	0.5226	0.4334

Table 2: Zero-shot prediction quality on the RO-EN dataset (Measured with Person's *r*).

6 Future Work

Building upon the previously discussed results, we propose ideas for future work. Considering the experimental design, the multitask architecture provides additional degrees of freedom that were not explored extensively, yet. For example, one could vary the amount of training per task or learn the training schedule as a parameter which adapts dynamically during the training process (Kiperwasser and Ballesteros, 2018; Zaremoodi et al., 2018). In addition, the number of auxiliary tasks could be increased to two or more, mixing different task types. To further evaluate the generalisability of the proposed methods, experiments with additional datasets, low-resource language pairs as well as alternative QE architectures and language models could be conducted, too.

Going beyond the field of Machine Translation Quality Estimation, it would be interesting to see the methods applied in adjacent areas of NLP. For example, this could entail closely related settings, such as quality estimation for machine-generated text summaries, as well as the fields of NLI and VQA, both of which face partial input biases. Other observable biases could also be considered as candidates for the use of targeted bias reduction techniques, provided that it is possible to design a counterbalancing auxiliary task or isolate the bias well enough to deploy adversarial approaches. We think that if the latter scenario applies, the adapted debiased focal loss technique for regression could be worth further exploration, too.

7 Conclusion

This paper expands on recent research which suggests that QE models are susceptible to learning spurious correlations. Based on additional analysis, and inspired by related work in the fields of NLI and VQA, we propose a range of auxiliary tasks that inform the main Quality Estimation task during training and are discarded at test time. First, we train the main Quality Estimation task together with additional, less biased data in a multitask setting. Then, we explore adversarial training and debiased focal loss to directly target the partial input bias. We find that the former approaches yield more stable results than the latter and conjecture that this is due to the difficulty of isolating partial input bias effects from useful predictive information encoded in the translation. We show that our proposed multitask architecture MultiTransQuest, especially when trained with additional shuffled WikiMatrix data, generalises well across the two most biased language pairs and the two different quality scores. Our method retains the overall prediction quality, reduces the observed biases significantly and increases the models' robustness in a zero-shot setting.

Acknowledgements

Marina Fomicheva and Lucia Specia were supported by funding from the Bergamot project (EU H2020 Grant No. 825303).

References

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. On Adversarial Removal of Hypothesis-only Bias in Natural Language Inference. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 256–262, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

Rich Caruana. 1997. Multitask Learning. *Machine Learning*, 28:41–75.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicolas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P. H. S., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N. Shamsi, Xudong Shen, Haoyue Shi, Yiwen Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, KV Aditya Srivatsa, Tony Sun, Mukund Varma T, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Gloria Wang, Zijie J. Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmański, Tianbao Xie, Usama Yaseen, M. Yee, Jing Zhang, and Yue Zhang. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation. arXiv.

Kai Fan, Bo Li, Feng-Ming Zhou, and Jiayi Wang. 2019. "Bilingual Expert" Can Find Translation Errors. In *AAAI Conference*. Association for the Advancement of Artificial Intelligence.

Marina Fomicheva, S Sun, E R Fonseca, Frédéric Blain, V Chaudhary, Francisco Guzman, N Lopatina, Lucia Specia, and A F T Martins. 2020. MLQE-PE : a multilingual quality estimation and post-editing dataset. arXiv.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In

Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, pages 1180–1189. JMLR.org.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-End Bias Mitigation by Modelling Biases in Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-Estimator. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 17(1).

Eliyahu Kiperwasser and Miguel Ballesteros. 2018. Scheduled Multi-Task Learning: From Syntax to Translation. *Transactions of the Association for Computational Linguistics*, 6:225–240.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Jan Niehues and Eunah Cho. 2017. Exploiting Linguistic Resources for Neural Machine Translation Using Multi-task Learning. In *Proceedings of the Second Conference on Machine Translation*, pages 80–89, Copenhagen, Denmark. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming Language Priors in Visual Question Answering with Adversarial Regularization. *Advances in Neural Information Processing Systems*, pages 1541–1551.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation Quality Estimation with Cross-lingual Transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sebastian Ruder. 2017. An Overview of Multi-Task Learning in Deep Neural Networks. arXiv.

Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, National University of Ireland, Galway.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F T Martins. 2020. Findings of the WMT 2020 Shared Task on Quality Estimation. In *Proceedings of the 5th Conference on Machine Translation (WMT)*, pages 743–764. Association for Computational Linguistics.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality Estimation for Machine Translation. *Synthesis Lectures on Human Language Technologies*, 11(1).

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8281–8291, Online. Association for Computational Linguistics.

Shuo Sun, Francisco Guzmán, and Lucia Specia. 2020. Are we Estimating or Guesstimating Translation Quality? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6262–6267.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba Submission for WMT18 Quality Estimation Task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 809–815, Belgium, Brussels. Association for Computational Linguistics.

Poorya Zaremoodi, Wray Buntine, and Gholamreza Haffari. 2018. Adaptive Knowledge Sharing in Multi-Task Learning: Improving Low-Resource Neural Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 656–661, Melbourne, Australia. Association for Computational Linguistics.

A Appendix

		DA (xlm-r-base)			HTER (xlm-r-base)		
		r	ρ	MSE	***r***	ρ	MSE
high-resource	**EN-DE**	**0.3695**±0.04	0.3874	0.0239	**0.4734**±0.01	0.4662	0.0308
	EN-ZH	**0.4249**±0.01	0.4155	0.0246	**0.3337**±0.01	0.3301	0.0792
mid-resource	**RO-EN**	**0.8467**±0.01	0.7914	0.0165	**0.7971**±0.01	0.6672	0.0416
	ET-EN	**0.6882**±0.01	0.7018	0.0520	**0.6695**±0.01	0.6646	0.0327
	RU-EN	**0.7133**±0.01	0.6781	0.0254	**0.3970**±0.01	0.3260	0.0613
low-resource	**NE-EN**	**0.7110**±0.01	0.6770	0.0184	**0.5462**±0.01	0.5313	0.0397
	SI-EN	**0.5880**±0.01	0.5427	0.0299	**0.5530**±0.04	0.5383	0.0393

Table 3: Pearson *r*, Spearman *p* and MSE for MonoTransQuest benchmark predictions on the test set (Direct Assessment & HTER) Note that we did our best to reproduce the results but reached a slightly worse performance. Possible reasons for the deviation are: the use of different random seeds, hardware or versions of the pre-trained XLM-R model.

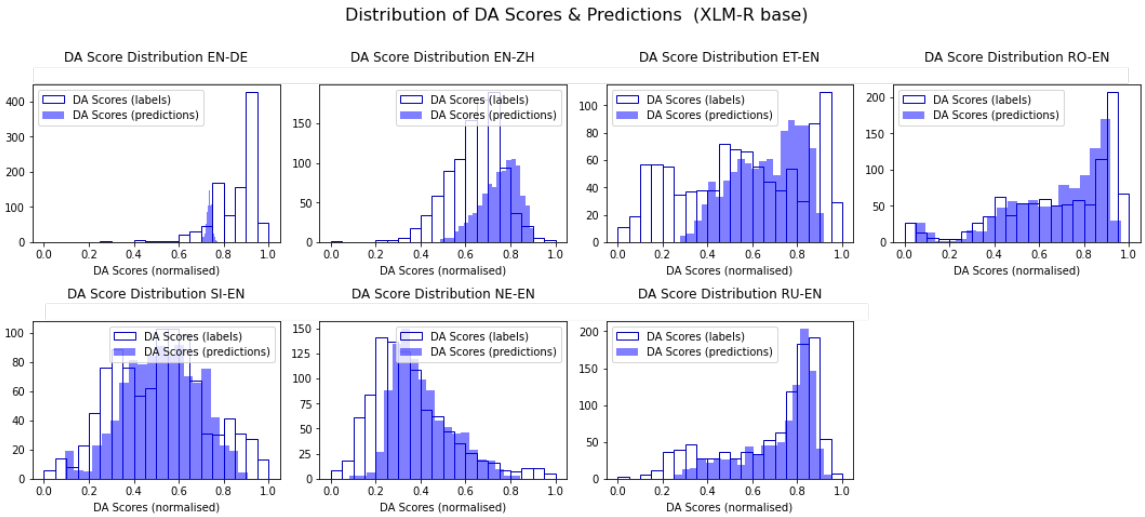


Figure 4: Distribution of MonoTransQuest DA predictions

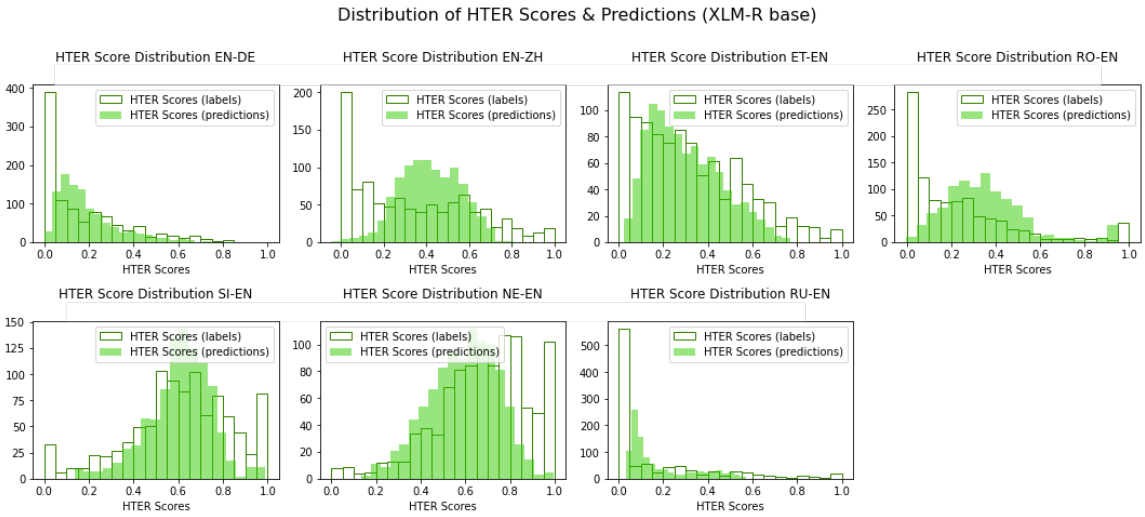


Figure 5: Distribution of MonoTransQuest HTER predictions

	Hyperparameter				DA				HTER			
#	Sep	Sha	LR Aux	Batch	*r*	ρ	MSE	*r (target)*	*r*	ρ	MSE	*r (target)*
1	1	0	2e-5	8	0.2839	0.3248	0.0160	0.2343	0.4819	0.4596	0.0304	0.4145
2	2	0	2e-5	8	0.3490	0.3723	0.0154	0.1967	0.4086	0.4161	0.0315	0.3726
3	3	0	2e-5	8	0.3630	0.4142	0.0141	0.1979	0.4542	0.4432	0.0303	0.4000
4	3	1	2e-5	8	0.3619	0.3781	0.0165	0.1808	0.4594	0.4450	0.0306	0.3812
5	3	0	3e-5	8	0.3578	0.3747	0.0167	0.2522	0.4459	0.4365	0.0330	0.3691
6	**3**	**0**	**2e-5**	**16**	**0.3811**	**0.4235**	**0.0175**	**0.1759**	**0.4630**	**0.4460**	**0.0300**	**0.4018**

Table 4: Approach 1a: EN-DE with RO-EN as auxiliary task and backpropagation per task *(Modified hyperparameter: Sep = Number of separate layers; Sha = Number of shared layers on top of XLM-R; LR Aux = Learning rate of the auxiliary task; Batch = Batch size)*

	Hyperparameter			DA				HTER			
#	Batch	Sep	Weight	*r*	ρ	MSE	*r (target)*	*r*	ρ	MSE	*r (target)*
1	16	3	50/50	0.3217	0.3304	0.0169	0.2768	0.4713	0.4645	0.0307	0.4118
2	8	3	50/50	0.3763	0.4115	0.0164	0.2915	0.4983	0.4794	0.0296	0.4254
3	**8**	**2**	**50/50**	**0.3625**	**0.3902**	**0.0163**	**0.2666**	**0.4956**	**0.4691**	**0.0297**	**0.4028**
4	**8**	**2**	**30/70**	**0.3314**	**0.3638**	**0.0165**	**0.0631**	**0.4698**	**0.4825**	**0.0307**	**0.3992**

Table 5: Approach 1b: EN-DE with RO-EN as auxiliary task and summed loss *(Modified hyperparameter: Batch = Batch size; Sep = Number of separate layers; Weight = Weighting of the tasks (main/auxiliary) in the loss function)*

	Hyperparameter			DA				HTER			
#	Batch	Sep	Augment	*r*	ρ	MSE	*r (target)*	*r*	ρ	MSE	*r (target)*
1	8	2	shuffle	0.2583	0.3423	0.0147	0.0026	0.4645	0.4169	0.0320	0.3222
2	8	3	shuffle	0.2357	0.4222	0.0179	-0.0349	0.4609	0.4378	0.0388	0.3133
3	**16**	**2**	**shuffle**	**0.4220**	**0.4431**	**0.0161**	**-0.0861**	**0.4489**	**0.4241**	**0.0412**	**0.1764**
4	16	3	shuffle	0.3481	0.3859	0.0172	-0.0521	0.4629	0.4386	0.0365	0.3560
5	16	2	context	0.2402	0.2891	0.0203	0.1206	0.4467	0.4304	0.0306	0.3345

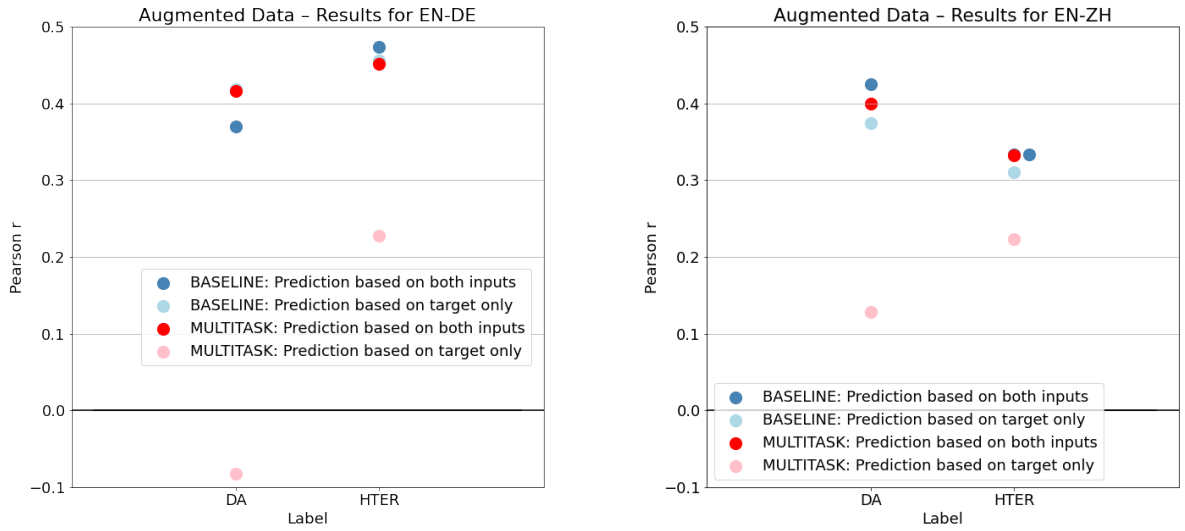
Table 6: Approach 2: Training with augmented WikiMatrix data as auxiliary task *(Modified hyperparameter: Batch = Batch size; Sep = Number of separate layers; Augment = Sentence augmentation strategy)*

	Hyperparameter			DA				HTER			
#	Batch	Sep	Grad Rev	*r*	ρ	*MSE*	*r (target)*	*r*	ρ	*MSE*	*r (target)*
1	16	1	-1	0.3015	0.3588	0.0184	-0.0868	0.4459	0.4075	0.0316	0.3221
2	16	2	-1	0.1738	0.2355	0.0231	0.0981	0.4619	0.4508	0.0332	-0.2574
3	16	3	-1	0.1160	0.2450	0.0172	0.0049	0.0921	0.1091	0.0374	-0.0744
4	**8**	**1**	**-1**	**0.3356**	**0.3957**	**0.0162**	**0.1049**	**0.4213**	**0.4089**	**0.0333**	**0.0548**
5	8	1	-0.5	0.3159	0.3714	0.0161	0.1084	0.4509	0.4357	0.0317	0.1153

Table 7: Approach 3: MultiTransQuest trained with target bias adversary *(Modified hyperparameter: Batch = Batch size; Sep = Number of separate layers; Grad Rev = Weight of the gradient reversal layer)*

	Hyperparameter		DA				HTER			
#	Batch	Weight	*r*	ρ	MSE	*r (target)*	*r*	ρ	MSE	*r (target)*
1	8	[0,1], $\beta = 1$	0.4380	0.4608	0.0144	0.4038	0.4648	0.4445	0.0326	0.4484
2	16	[0,1], $\beta = 1$	0.4027	0.4289	0.0148	0.3574	0.4623	0.4453	0.0306	0.4010
3	16	[0,1], $\beta = 2$	0.4112	0.4313	0.0146	0.3470	0.4363	0.4193	0.0313	0.3822
4	**16**	**[0,1],** $\beta = 3$	**0.4104**	**0.4189**	**0.0152**	**0.3320**	**0.4713**	**0.4530**	**0.0302**	**0.4022**
5	16	[0,1], $\beta = 3.5$	0.3394	0.3745	0.0158	0.2764	0.4462	0.4442	0.0320	0.3843
6	16	[0,1], $\beta = 4$	0.3323	0.3391	0.0155	0.2885	0.4472	0.4478	0.0324	0.4119
7	16	[0,1], $\beta = 3$ S-shaped	0.3322	0.3580	0.0155	0.2913	0.4661	0.4365	0.0299	0.3961

Table 8: Approach 4: MonoTransQuest model trained with target bias focal loss *(Modified hyperparameter: Batch = Batch size; Weight = Weighting of the bias model)*



(a) EN-DE with shuffled data as aux task

(b) EN-ZH with shuffled data as aux task

Figure 6: Shuffled WikiMatrix data: Performance and partial input bias reduction

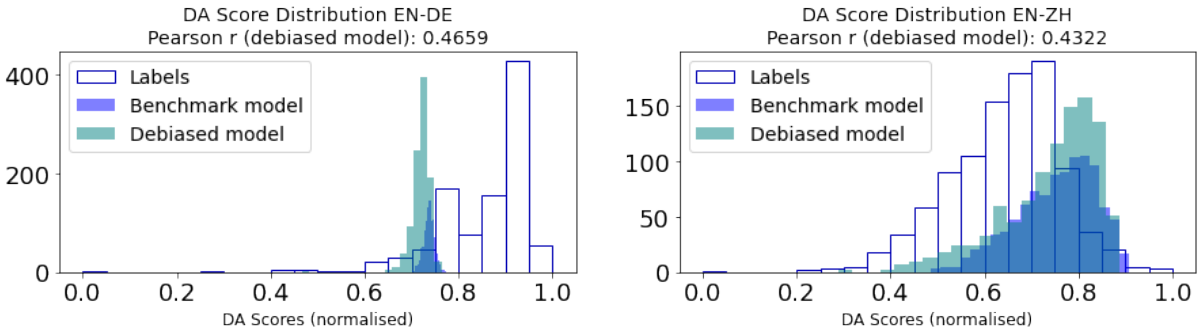


Figure 7: MultiTransQuest: DA prediction distribution

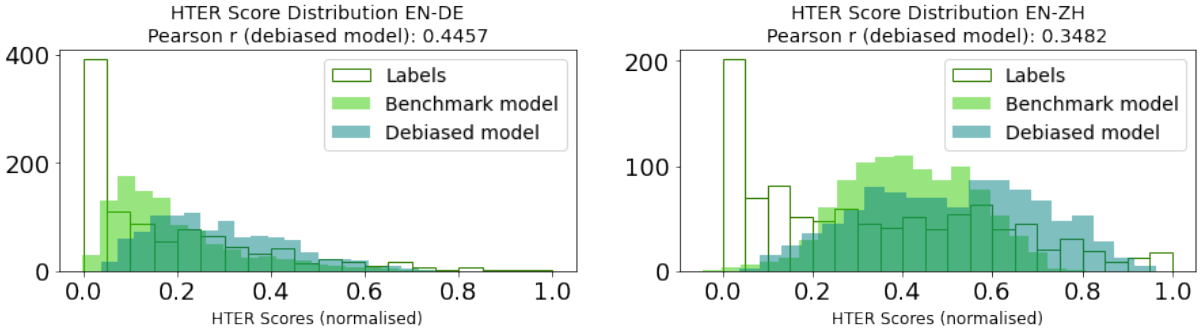


Figure 8: MultiTransQuest: HTER prediction distribution