PreQuEL: Quality Estimation of Machine Translation Outputs in Advance

Shachar Don-Yehiya **Leshem Choshen** **Omri Abend**
School of Computer Science and Engineering, The Hebrew University of Jerusalem
{first.last}@mail.huji.ac.il

Abstract

We present the task of *PreQuEL*, Pre-(Quality-Estimation) Learning. A PreQuEL system predicts how well a given sentence will be translated, without recourse to the actual translation, thus eschewing unnecessary resource allocation when translation quality is bound to be low. PreQuEL can be defined relative to a given MT system (e.g., some industry service) or generally relative to the state-of-the-art. From a theoretical perspective, PreQuEL places the focus on the source text, tracing properties, possibly linguistic features, that make a sentence harder to machine translate.

We develop a baseline model for the task and analyze its performance. We also develop a data augmentation method (from parallel corpora), that improves results substantially. We show that this augmentation method can improve the performance of the Quality-Estimation task as well.[1] We investigate the properties of the input text that our model is sensitive to, by testing it on challenge sets and different languages. We conclude that it is aware of syntactic and semantic distinctions, and correlates and even over-emphasizes the importance of standard NLP features.

1 Introduction

Can we tell if a sentence is difficult to automatically translate, without actually translating it? We argue that this question has important practical and theoretical value.

From a practical standpoint, a PreQuEL system can save trouble and cost. For individuals or companies who do not maintain their in-house Machine Translation (MT) system, translating a big amount of data might turn out to be expensive. There are several automatic translation services offered, but there is no prior indication that any of them would do a good enough job. Therefore, after purchasing translations, one would probably want to estimate the quality of the translations, with human annotators or an automatic Quality Estimation (QE) system (Blatz et al., 2004; Specia et al., 2009). After spending these resources, there is still a chance that the translation quality will be judged unsatisfying, and there will be no choice but to hire human translators. The decision between the automatic route and the much more expensive alternative of hiring human translators can be informed by a PreQuEL model, before investing the resources to actually translate. See figure 1.

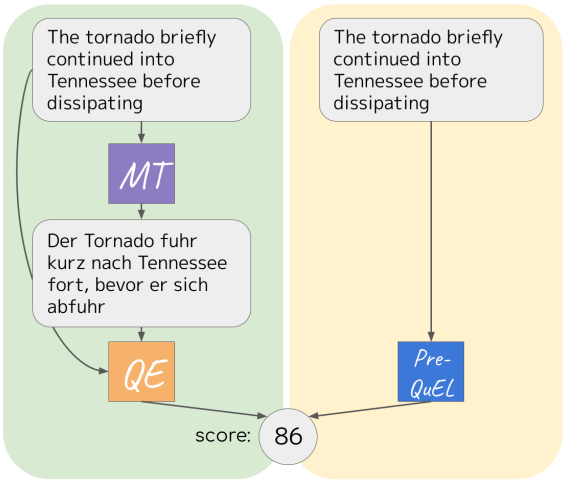

Figure 1: Quality Estimation (QE) vs. PreQuEL. In the PreQuEL task, the score is generated from the source sentence directly, before investing the necessary resources to translate.

Theoretically, PreQuEL allows insight into the performance boundaries of state-of-the-art MT methods. It is common practice to search in retrospect what kinds of text phenomena are difficult for existing MT systems to handle (Toral and Sánchez-Cartagena, 2017). Such analysis helps in finding flaws in the system, and getting some notion of what is needed to overcome them. Challenge sets are also used for that purpose – testing the system, analyzing its performance on carefully chosen

[1] Code and data are available in: https://github.com/shachardon/PreQuEL

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11170 - 11183
December 7-11, 2022

cases (Isabelle et al., 2017; Barrault et al., 2020; Choshen and Abend, 2019). However, no attention was heeded to the source sentence itself and how it affects the translation quality. Are there any specific properties, maybe linguistic features, that make a sentence more difficult for MT? By introducing the PreQuEL task, we place the focus on the input text, and learn which source texts are easy to translate and which are difficult.

We consider two variants of this approach. One, where the system is given and another where the prediction is done relative to the state-of-the-art in the field. The latter of course bears a tacit assumption that there are underlying properties that make a sentence easier or harder to translate for a state-of-the-art system. Supporting empirical evidence for this assumption is presented in §7.1.

We develop a baseline model for the task, based on a state-of-the-art QE system, and report its results in §6. We further develop an automatic data augmentation method from parallel corpora and use it in intertraining (Phang et al., 2018) and multi-task settings, outperforming our initial results (§3.2). In §9, we show that this augmentation method can improve the performance of the QE task as well. Our results, surprisingly, do not fall far behind the results of state-of-the-art QE systems, despite not being exposed to the actual translation.

To the best of our knowledge, this work is the first to address PreQuEL. Sun et al. (2020) did point out that it is possible to perform QE using only the input sentence, but viewed this finding as an artifact of their dataset. We revisit their claims and argue that the PreQuEL model can potentially simulate the MT system, and there is thus no theoretical reason to consider the dataset as "cheatable". We discuss their claims extensively in §9.

Analyzing the proposed model, we examine in §7.1 the question of whether the predictions made by the model are specific to the performance of one particular system (whose outputs were used for supervision), or whether they generalize to other recent neural systems. Our results indicate that results do generalize to other systems, finding a drop of correlation of only 4.2 points.

We further examine the contribution of the syntactic structure of the output in predicting its difficulty. To do so we experiment with a model that was fine-tuned with syntactic parsing, but find that its improvement to the performance is relatively low (§4.2). However, we do find a correlation between the baseline model's predictions and some syntactic features, suggesting that the model is aware of syntax (§7.2). We further explore what features of the input sentence our model is sensitive to, and report results in §7.5.

2 Task Definition

PreQuEL is the task of predicting the likelihood of an MT system to correctly translate a given sentence to a given target language. The task has two variants.

The system-specific variant is given a source language S, a target language T and an MT system between them M. The goal is to learn a function $g : S \rightarrow \mathbb{R}$, such that for every sentence $s \in S$ the score $g(s)$ represents the expected quality of the translated sentence $M(s) = t \in T$. As standard in QE evaluation, the PreQuEL predictions are evaluated against the gold labels using Pearson's r correlation coefficient (Specia et al., 2020).

The definition for the second variant, namely of general state-of-the-art MT, is equivalent but with no specific M. Instead, g satisfies the above for all M such that M is a state-of-the-art MT. Unlike the first, the correlation achievable in this variant is inherently bound, due to the non-uniform behavior exhibited by different MT systems. Nevertheless, to the extent that systems do share performance patterns, mapping them is of theoretical and applicative value.

We focus on the first variant at first and examine the second in §7.1.

3 Data

3.1 Quality Estimation Data

The WMT shared task on QE includes estimation at three granularity levels: word, sentence, and document. The WMT 2020 (Specia et al., 2020) introduced a variant of the sentence-level task. The sentences were annotated with direct assessment (DA) scores (Graham et al., 2013), instead of labels based on post-editing (Snover et al., 2006). The dataset for the task is composed of data extracted from Wikipedia for six language pairs, of which we use the high-resource languages English-German (en-de) and English-Chinese (en-zh) and medium-resource pair Estonian-English (et-en). Each language pair has 7K,1K,1K sentence pairs in the training, development and test sets respectively. Translations were produced with a state-of-the-art MT model built using the fairseq toolkit (Ott et al.,

2019). Each translation was rated following the FLORES guidelines (Guzmán et al., 2019). For our purposes, we take only the source sentences and the DA scores, ignoring the translations. To facilitate training, we transform the scores to a 0-1 scale with a min-max normalization.

3.2 Augmented Data

We propose an automatic method to acquire more translation quality scores, dispensing with the costly human annotations.

Automatic evaluation metrics (Mathur et al., 2020) give additional information on translation quality. Unlike QE, such metrics are exposed to human translation as well. While relying on QE for augmentation would possibly be beneficial, we opt for extracting the latent information found in automatic metrics instead.

Given a parallel corpus, we re-translate it with the MT system M. We take the source sentences to be the PreQuEL inputs, and the metric scores to be the labels.

Where automatic metrics are often far from agreeing with human judgments (Choshen and Abend, 2018; Kocmi et al., 2021), they still extract some potentially beneficial aspects of similarity to the translation. As the PreQuEL has no access to those references (or translations), training to predict the automatically estimated quality of the augmented data encourages the PreQuEL system to extract those features.

We examine the difference between predicting DA and metric scores directly in appendix §C.1. Nonetheless, we show it is similar enough to be beneficial as a data augmentation method in §6.

4 Models

4.1 Baseline Model

We propose a baseline system model for PreQuEL. Our implementation builds on TransQuest (Ranasinghe et al., 2020), the winner system of the WMT 2020 QE sentence-level DA task (Specia et al., 2020). The model is builds on RoBERTa-large (Liu et al., 2019) to derive the representations of the input sentence. For pooling it uses the output of the [CLS] token. After pooling we place a 1024×1024 and 1024×1 fully-connected layers. The latter produces the final output.

4.2 Advanced Models

We examine the advantages of more advanced architectures. These architectures combine syntactic or semantics components, that may improve PreQuEL results. Additionally to that, our other motivation is to get a better understating of PreQuEL required features. In the case of a significant improvement over the baseline models, we would conclude that syntactic and semantic knowledge are crucial for PreQuEL.

COMBINED. To examine the need for external syntactic knowledge, we combine a UD parser with our architecture. We use RobertNLP (Grünewald and Friedrich, 2020), which is also RoBERTa based. Using the parser as an intertraining step resulted with lower performance. Instead, we concatenate the last hidden layer of a fresh pre-trained RoBERTa model and that of the parser. The concatenated hidden layers are used as the input to the classifier (same as the one in the original architecture, with adjustments to the dimensions).

MULTITASK. Under certain circumstances, related tasks may help each other using Multi-task Learning (MTL) (Aghajanyan et al., 2021). As discussed in §3.2, the task of predicting other automatic metrics might help predict DA too. We extend the model with additional classifier heads, each predicting a different automatic metric.

5 Experimental Setup

We train and test our PreQuEL models each time on one language pair. The first pair we examine is the high-resource en-de. As discussed in Fomicheva et al. (2021), the translation quality for translations in this pair has little variability, with a mean score of 73.25 and std 8.13. The vast majority of translations were assigned high DA scores, which makes differentiating between them challenging.

To balance this, we select as our second pair the medium-resource et-en. Non-high-resource pairs give QE models an advantage – outputs are occasionally 'hallucinated', i.e., they do not have anything to do with the original sentences. Detecting such cases should be simple for QE systems, which explains the high QE scores on those pairs. We would therefore like to examine whether this effect persists in the PreQuEL settings, where the outputs are not available to the model.

We use the pipeline described in §3.2 to create scores for en-de and en-zh (when avail-

able) WMT-News (Barrault et al., 2020), bible-uedin (Christodoulopoulos and Steedman, 2015), Tatoeba (Tiedemann, 2020) and GlobalVoices (Tiedemann, 2012), all taken from OPUS (Tiedemann, 2012). We remove duplicate sentences and randomly split each of the datasets to train/dev/test, 80%/10%/10%. For the MT system we use the OPUS-MT released model (Tiedemann and Thottingal, 2020), based on Marian-MT (Junczys-Dowmunt et al., 2018).

The selected datasets are diverse in their domain. WMT-News (28,887 sentences) is a parallel corpus of news provided by WMT for testing MT performance. bible-uedin (48,705 sentences) is a multilingual parallel corpus created from translations of the Bible. Tatoeba (197,381 sentences) is a crowd-sourced collection of user-provided translations. GlobalVoices (55,822 sentences) is a news parallel corpus from the website Global Voices. We examine the out-of-domain effect in App. §C.2. We also ensure that for each language pair at least two datasets were not found in Marian training set, making PreQuEL predictions more interesting.

We note that despite we translate from English to German/Chinese, the translation direction is not always from English. This might lead to artifacts in the metric evaluation (Graham et al., 2020). However, we speculate this is more of a concern for evaluation. Most of our experiments rely on the data as an augmentation method, which either helps the main task or not.

For the automatic metric we use COMET (Rei et al., 2020). Where more than one metric is required (for training MULTITASK), we use also ChrF++ (Popović, 2015) and BERTScore (Zhang* et al., 2020). These three metrics cover different kinds: ChrF++ is string-based, BERTScore is an unsupervised embedding-based model, and COMET is a supervised model trained end to end.

We train instances of SIMPLE for en-de and et-en. For en-de and en-zh we train also a SIMPLE Aug version with COMET as intertraining. For the en-de en-zh comparison (§7.4), we use datasets that are available in both languages: NewsTests and bible-uedin, and test them on the subset of source sentences that are shared between en-de and en-zh development DA. For en-de we also train the more advanced architectures, COMBINED and MULTITASK. We train two versions of COMBINED, COMBINED+ and COMBINED- , the first concatenated to an actual UD parser and the latter to another pre-trained RoBERTa to control for the size. We use the augmentation intertraining for both versions. Where we compare two or more versions, we train 3 seeds for each.

We take the hyperparameters from the TransQuest implementation, Adam optimizer with a learning rate $1e-5$, and a linear learning rate warm-up over 10% of the training data. We adjust the batch-size from 8 to 4 to fit our GPUs. See App. A.

To allow evaluation during training, we sampled 10% of the training data and kept it for evaluation. This is in addition to the development data we use for evaluation at the end of the training. We carry out evaluation every 300 training steps for a small training set (smaller than 1K steps), and every 3K steps otherwise. We perform early stopping with patience of 10 evaluation rounds. The model is trained for a maximum of 3 epochs, and no less than one. If we early-stop during the first epoch (Dodge et al., 2020), or reach a low correlation (< 0.1) on the last evaluation round, we reset the seed controlling initialization, batches, split of the training/evaluation data, etc. In order to improve our results, we train 3 random seeds and infer by an average ensemble. We carry out evaluation on the DA test set.

Since this is the first work on PreQuEL, there are no immediate candidates for baselines. An exception is discussed in §9.2 and App. §C.4. We take the negated length of the sentences as a baseline. We assume the longer a sentence is, the harder it is to translate. We experimented also with the perplexity score of GPT-2 (for English as the source language), resulting with low correlation (0.06). We discuss more correlated features in §7.5. In places where we compare ourselves to a QE system, we use TransQuest. Also here we adjust the batch-size to 4 to fit our GPUs.

6 Main Results

In Table 1, we present the main results. All of our models outperform the baseline. Similar to the case in QE, en-de is more challenging than et-en. The augmentation improves the results in both settings – intertraining (SIMPLE Aug) and multi-task (MULTITASK), confirming the value of our augmentation method. MULTITASK outperforms the other architecture, with an improvement of more than 14 points over SIMPLE, suggesting that different metrics capture different properties that can be use-

	en-de	et-en
PreQuEL Models		
SIMPLE	0.196±0.02	0.602±0.00
SIMPLE Aug	0.315±0.00	-
COMBINED+	0.326±0.02	-
COMBINED-	0.265±0.00	-
MULTITASK	**0.336±0.01**	-
Baseline	0.135	0.050
Upper Bound (QE)		
TransQuest	0.381±0.04	0.767±0.00

Table 1: Pearson's r of the predictions of the PreQuEL models with the DA scores, for en-de and et-en. All our models outperform the baseline, and the augmentation improves the results both as intertraining and multitask. The et-en pair correlation is much higher than the correlation of en-de.

ful for PreQuEL. As for COMBINED, COMBINED+ (with parser) outperforms COMBINED- (same size model, without parser) by more than 6 points, confirming the value of the syntactic knowledge, and not just the model size. COMBINED+ outperforms SIMPLE Aug by 1 point. However, this improvement was achieved at the cost of doubling the size of the model, and is outperformed by the much smaller MULTITASK.

We use TransQuest as our upper-bound, to compare the performance of a similar QE model. MULTITASK is outperformed by TransQuest[2] by 4.5 points.

7 Analysis and Discussion

7.1 General state-of-the-art MT Systems

So far, we showed results for the first variant of the task, namely for predicting the quality of outputs that were created using a single known MT system. Here we approach the second variant, the one that predicts the score for a general state-of-the-art MT system. We expect a PreQuEL model for this variant to predict the quality score for any state-of-the-art MT system, without any supervision of it. Our experiments with this variant confirm our assumption. Some shared properties make a sentence easier or harder to translate, across systems.

Manual DA annotations are scarce, and there is no data for other systems except for the WMT's DA of the MT task participants (Barrault et al., 2020), which is partial and small. Given the correlation between DA and reference-based metrics, we approximate the data using the augmentation pipeline described in section 3.2, this time translating with Facebook FAIR (Ng et al., 2019). The correlation of the train COMET labels of Marian with the train COMET labels of Facebook FAIR ranges from 0.60 to 0.82, depending on the dataset. This result implies that although there are differences in the translation quality, there is a lot in common too. To test what of this similarity our model catches, we take a SIMPLE that was trained on data created with Marian and test it on data created with Facebook FAIR.

The correlation on the Marian test set is 0.652, while the correlation on the Facebook FAIR test set is 0.610. One might expect the PreQuEL model to be bounded by the correlation between MT outputs ($0.82 \cdot 0.652 = 0.535$). However, these results do not fit this hypothesis ($0.610 > 0.535$) and indicate that the PreQuEL model generalizes well to other systems. Everything above the point of similarity in MT systems predictions is evidence of the preference of PreQuEL towards shared cues for MT system performance.

7.2 Word-Ordering

To further investigate the role of syntax, we use two German word-ordering challenge sets (Choshen and Abend, 2021). These two datasets [3] consist of pairs of sentences, each pair consists of two sentences, both holding the same meaning. They differ only in the order of the subject and object, the first is in the more common order, subject before object, and the second is object before subject.

On the first dataset, it is the case that resolves the ambiguity and determines which is the subject and who is the object. For example, *"Das Kind bringt den Ball"* and *"Den Ball bringt das Kind"* should be both translated to *"The child brings the ball"*.

On the second dataset, it is the verb that determines this. For example, *"Die Katzen kicken die Maschine"* and *"Die Maschine kicken die Katzen"* should be both translated to *"The cats kick the machine"*. Additionally, on the second dataset each pair has its reverse-pair, in which the subject and object switch their roles. So we also have *"Die Maschine kickt die Katzen"* and *"Die Katzen kickt die Maschine"*, where both should be translated to *"The machine kicks the cats"*.

Similar to §7.1, we approximate the data with the augmentation to allow data for training de-en SIMPLE. We use it to predict the scores of the

[2]Despite using the official code and our best efforts, TransQuest performance is lower than reported in its paper.

[3]Challenge sets are available in: https://github.com/borgr/nematus

dataset. We measure the correlation between the predictions of the four versions we have for each sentence in the second dataset: 1.subject-object, 2.object-subject, 3.reversed-meaning-subject-object, 4.reversed-meaning-object-subject. The correlation between sentences with the same meaning (1-2 and 3-4) is higher than the correlation between sentences with the same syntax (1-3 and 2-4) or sentences that are more similar in terms of their linear word order (1-4 and 2-3). For the full results see App. §D. We conclude that meaning plays a more important role for our model than the syntax and the linear word order.

We compare the mean predictions of the subject-object order with the object-subject order. In the first dataset the mean score for the more common order subject-object is higher. In the second dataset however, for the reversed pair subject-object is indeed higher, but for the non-reversed pair the object-subject is the higher (See App. §D).

7.3 The Model as an Analytic Method

Following the results of the word-ordering challenge set, we consider the possibility of using the model as an analytic method. Concretely, we compare the PreQuEL scores for the source sentence, and a modified version of it (e.g., in past tense). This allows inspection of the effect linguistic aspects of the source have on translation quality.

We conduct our analysis on the following transformations proposed and implemented by NL-Augmenter repository (Dhole et al., 2021): *GenderSwap*, *TenseTransformation*, *RandomDeletion*, *YesNoQuestionPerturbation*, *ChangePersonNamedEntities*, *MultilingualBackTranslation*, *ReplaceNumericalValues*, *YodaPerturbation*.[4] See App. §E.

Transformations might not affect some sentences. In the *TenseTransformation-past* for example, if the sentence is already in the past tense, it shouldn't change. Therefore, we report the number of sentences that were changed, and the mean score before and after the transformation for this subset of sentences. For prediction, we use en-de SIMPLE Aug.

The transformations we use are automatic. Hence, some level of noise is expected, which may lower scores. Nevertheless, Table 2 shows that there are cases where the difference between the original sentences and the transformations is relatively small, and cases where it is more substantial. For example, *ChangePersonNamedEntities* and *ReplaceNumericalValues* show a difference smaller than 0.01, in agreement with our expectations. That is, we would not expect a change of a personal name or a numerical value to affect the performance of an MT system. On the Other hand, *RandomDeletion* and *YodaPerturbation* show a bigger difference, in agreement with the more drastic changes they make.

Examining *TenseTransformation*, we see that the past and present are similar, but the future is perceived to be harder to translate. This might be due to the future tense being less common.

Previous work suggested that translationese renders a text simpler (Baker et al., 1993), or at least different (Rabinovich and Wintner, 2015) than the source text. Table 2 shows that when translating the sentences to German and then back to English, their mean score is indeed higher, rendering back-translation easier to translate.

7.4 Different Target Languages

The syntactic and semantic structures converge and diverge between different languages, and there are therefore cases in which the translation of one language into another results in a very different structure than that of the source (Nikolaev et al., 2020;

[4]The computationally demanding nature of these experiments prohibit us from using the entire set of transformations.

Transformation	# sent	src	trans	diff
GenderSwap	324	73.72	73.64	-0.08
RandomDeletion	975	73.77	73.15	**-0.62**
YesNoQuestion	553	73.83	73.51	**-0.32**
ChangePersonName	148	73.78	73.74	-0.03
BackTranslation	996	73.78	73.89	**+0.11**
ReplaceNumericalVals	127	73.54	73.51	-0.03
YodaPerturbation	955	73.79	72.94	**-0.85**
Tense-past	403	73.78	73.64	**-0.16**
Tense-present	752	73.73	73.65	-0.08
Tense-future	904	73.76	73.50	**-0.26**

Table 2: Mean predictions of a SIMPLE Aug before and after applying transformations. Differences greater than 0.1 are. boldfaced.

Model	en-de DA	en-zh DA
SIMPLE en-de	**0.377**	0.260
SIMPLE en-zh	0.140	**0.577**

Table 3: Pearson's r of the predictions of two SIMPLEs Aug: one trained on en-de data and one on en-zh, with the matching DA scores for both en-de and en-zh. Both models have a higher correlation with the gold labels of the development set of the language that they were trained on.

	length	depth	LM	VERB	advcl	case	unigram	bigram	3gram	4gram
model preds	**-0.2894**	**-0.2157**	**0.2151**	**-0.3144**	**-0.2541**	**-0.2056**	**0.2230**	**0.3258**	**0.3362**	**0.3338**
gold labels	-0.1305	-0.0338	0.0968	-0.1424	-0.0848	-0.0822	0.1045	0.1546	0.1638	0.1625

Table 4: Pearson's r of SIMPLE Aug predictions and the gold labels of the en-de DA dev, with the the selected features. The correlation with the features is higher for the model predictions than for the gold labels, implying that our model overestimates the importance of the features.

Dorr, 1994). Does PreQuEL capture hard sentences in the seen source, or does it implicitly capture divergences between the source and target languages?

To assess this we consider en-de and en-zh pairs. Like en-de, en-zh is both high resourced and obtains high QE results (Specia et al., 2020). The correlation of their DA scores is 0.08, meaning that despite having the same input sentences, what is considered hard is quite different.

In the PreQuEL settings on the other hand, the model is exposed only to the source sentences. The target sentences are only indirectly exposed by the DA score. Would it be enough for the model to learn what is hard to translate to one language, but not to another?

We use instances of SIMPLE Aug for en-de and en-zh to predict the scores for both en-de and en-zh development DA. We compare the correlation of the models' predictions with each other and the correlation of each model with the gold scores of each target language.

In Table 3, we see that both models have a higher correlation with the gold labels of the development set of the language that they were trained on. We conclude that the models manage to learn a function that is specific not only to the source language but also to the target language. The model predicts the difficulty of translating a sentence from a specific language to a specific language, and not just the general difficulty of the sentence. However, the correlation between the predictions of the models with each other is 0.588, much higher than the correlation between the DA labels. Thus, compared to the DA labels, the PreQuEL model does tend to overestimate the source.

7.5 Correlating with Standard Features

We further examine the correlation between our model predictions and standard NLP features. We consider sentence length, Universal Dependencies (UD) (Nivre et al., 2020) parse tree depth, number of edges with a specific UD label, number of tokens with a specific POS label, language detector confidence, language model score, and n-gram model probabilities. See App. B.

We report only the features that are statistically significant ($P < 0.0006$) after Bonferroni correction and that show at least 0.2 correlation with the predictions or the labels – sentence length, tree depth, language model score, verb POS count, adverbial clause label count, case label count, and n-grams.

In Table 4, we can see that the correlation with the features is higher for the model predictions than for the gold labels. This implies that our model not only uses these features, but also over-estimates their importance.

8 Determining whether to MT

We expect that the standard test case for PreQuEL would be to determine the chances of a sentence to be translated correctly to its target language. According to the FLORES guidelines, a translation that closely preserves the semantics of the source sentence gets a DA score of 70-90. Only a perfect translation gets a score of 91-100. Therefore, we report the Precision/Recall curve for a threshold of 70, to find the 'good enough' sentences, following (Guzmán et al., 2019). In App. F we also report a threshold of 90 to find the perfect ones.

We use the predictions of SIMPLES Aug to estimate what sentences will get a DA score above 70. Figure 2 shows the Precision/Recall curves for en-de and et-en, for a DA threshold of 70. For comparison, we plot the curves for a random model that uniformly predicts a random score between 0-100 for each sentence. For en-de, 76% of the sentences have a DA score above 70, so we can maintain a relatively high precision simply by assigning random scores. Still, our model outperforms it. For et-en, 43% of the sentences are above 70.

9 Quality Estimation

9.1 Augmentations for Quality Estimation

As mentioned before, our data augmentation method might be useful for training QE systems too. QE still has no access to the reference that metrics rely on. Extracting similarities between the

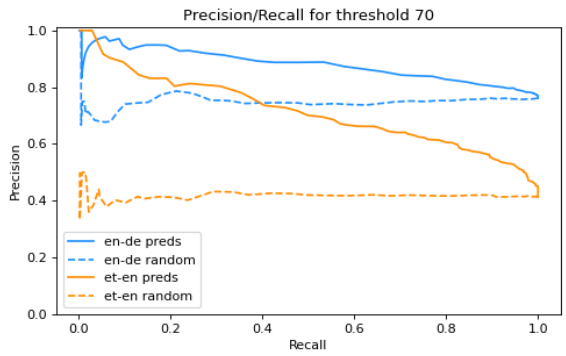

Figure 2: Precision/Recall curve for en-de and et-en development datasets, for the prediction of whether a DA threshold of over 70 would be given to the translation of the input sentence. The blue plots are comparable and so are the orange (but not cross-color comparisons).

translation to a hidden reference might force new abilities to emerge in the QE networks.

We train TransQuest on en-de DA data, with and without intertraining on COMET data. We run each version 3 times with different seeds, to confirm the results. The average correlation for TransQuest without the augmentations is 0.381 with std 0.043 The average correlation for TransQuest with the augmentations is 0.429 with std 0.008. TransQuest thus benefits from the augmented data, in terms of both correlation and stability (much smaller std).

9.2 Quality Estimation Evaluation

Sun et al. (2020) argue that the recent success of pre-trained language models for QE is overstated. They point out that it is possible to perform QE to a large extent using only the source or output sentence (we outperform their results; see App. §C.4). They viewed this finding as an artifact of the dataset, as they expect the predictions of a well QE system to reflect both the translation's closeness to the source text, and how well it fits in the target language. The good results the model obtained with only source/translation imply that only one of the two is taken into account. They suggest to replace HTER with a metric that represents both fluency and adequacy, such as DA.

We interpret these results differently. Indeed, these experiments provide motivation for PreQuEL, showing that the source sentence holds considerable information required for predicting system performance. However, we argue that an oracle PreQuEL model can simulate the MT system, and therefore there is no theoretical reason to consider the dataset as "cheatable".

As for their suggestion to use DA, we find that their criticism is not specific to HTER. We train a SIMPLE on en-de HTER data and compare its performance to the performance of a SIMPLE that was trained on the en-de DA data. We run each ensemble 3 times with different seeds, to confirm the results. The DA model gets an average correlation of 0.196 with std of 0.024. The HTER model gets an average correlation of 0.322 and std 0.013. These results suggest that predicting the DA score from the source alone is indeed somewhat more difficult than predicting the HTER, but definitely possible.

However, we agree that QE systems should not ignore parts of their input, and that this should be addressed in evaluation. QE systems in the WMT shared task are trained and evaluated against the outputs of a single MT system. Possibly, the QE systems learn to simulate it, which would explain the redundancy of the translation.

Instead, we suggest to train and test the QE models on datasets of multiple MT systems. Such a dataset would include translations from diverse systems. This would sever the ability of the QE systems to simulate the translation process implicitly. Therefore, QE systems would have to rely on the actual translation for successful prediction.

Moreover, to assert reliance on the source, translation of unrelated sentences should be included in the translation as well.

10 Conclusion

We presented a new task, PreQuEL, the task of predicting the quality of the output of MT systems based on the source sentence only. We developed a baseline model and reported its results, providing motivation for the task by showing that considerable information for predicting the quality scores is stored in the source sentence. We developed an automatic augmentation method, and used it to improve our results. We showed that the predictions made by a model that was trained with a specific system supervision do generalize to other state-of-the-art systems. We analyzed our model by testing it on challenge sets and other languages, concluding that our model is aware of syntax, meaning, and the target language.

Motivated by these latter results, we suggested to use the PreQuEL model as an analytic method, to confirm the effect of linguistic and semantic phenomena on the ability of a MT system to translate.

Future work will use this method to provide insights into the performance boundaries of current MT systems. Other lines of work we intend to pursue include examining the advantages of a multilingual PreQuEL model, and developing an advanced PreQuEL model that selects the MT system that is most likely to generate the best translation for a given text.

11 Limitations

Although our PreQuEL model for the en-de language pair reaches good results when compared to the TransQuest upper bound, the correlation is still not high in absolute terms (0.336 §6). Therefore, its predictions should be used with care.

One of our motivations in this paper is to investigate whether there are any linguistic features, that make a sentence more difficult for MT. We trained models for the task and reported their performance, supporting the claim that there are such features. However, trained with an end-to-end approach, our models are not suitable for explicitly pointing out these features. That is, when our models predict a sentence to be hard to translate, we can not tell why. We did however show some properties that our models are sensitive to. For example, the COMBINED architecture (§4.2) confirmed the role of syntax, and the German word ordering challenge sets (§7.2) revealed the importance of meaning. We also showed a correlation with standard features, to demonstrate their influence.

Another limitation is the available data. In some of the analysis experiments we needed data for languages pairs that do not appear in the shared task on QE (Specia et al., 2021), or translation outputs coming from multiple MT systems. In these cases, we used automatically augmented data (§3.2), instead of proper DA labels. This data is naturally of lower quality. We note that this limitation refers to some of the analysis only, and not our main results, which we tested on DA data only.

Acknowledgments

We thank Anna Pellivert and Menachem Shefer for helpful discussions. This work was supported in part by the Israel Science Foundation (grant no. 2424/21), and by the Applied Research in Academia Program of the Israel Innovation Authority.

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mona Baker, Gill Francis, and Elena Tognini-Bonelli. 1993. *'Corpus Linguistics and Translation Studies: Implications and Applications'*, pages 233–248. John Benjamins Publishing Company, Netherlands.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. 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.

Leshem Choshen and Omri Abend. 2018. Inherent biases in reference-based evaluation for grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Melbourne, Australia. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 2019. Automatically extracting challenge sets for non-local phenomena in neural machine translation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 291–303, Hong Kong, China. Association for Computational Linguistics.

Leshem Choshen and Omri Abend. 2021. Enhancing the transformer decoder with transition-based syntax. *ArXiv*, abs/2101.12640.

Christos Christodoulopoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49:375 – 395.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised

cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv*, abs/2002.06305.

Bonnie J. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2021. Mlqe-pe: A multilingual quality estimation and post-editing dataset.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Stefan Grünewald and Annemarie Friedrich. 2020. RobertNLP at the IWPT 2020 shared task: Surprisingly simple enhanced UD parsing for English. In *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 245–252, Online. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend. 2020. Fine-grained analysis of cross-linguistic syntactic divergences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1159–1176, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *ArXiv*, abs/1811.01088.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. Transquest at wmt2020: Sentence-level direct assessment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1049–1055, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, 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 Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

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, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight Interna-*

tional Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA).

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Computing Infrastructure

Our architecture is roberta-large based, and therefore the number of parameters for each instance of our model is 355M×3 due to ensembling. The COMBINED architecture uses two instances of roberta-large, and therefore is 355M×3×2.

When training a PreQuEL model with a source language that is not English, we replace roberta-large with xlm-roberta-large (Conneau et al., 2019).

We train each instance of our PreQuEL models on 2 CPU and 1 GPU. The run-time was 2.5 hours for a training set of 7K sentences. The COMBINED architecture with intertraining took the longer to train, 55 hours.

B Extracting Standard NLP Features

To extract the language detector confidence we use spacy. To extract the n-grams we use kenlm (https://github.com/kpu/kenlm). To extract POS tags, UD labels and UD tree depth we use stanza (Qi et al., 2020). To extract language model score we use lm-scorer (https://github.com/simonepri/lm-scorer).

C Additional Experiments

C.1 DA vs. COMET

To examine the differences between the DA and COMET data, we want to conduct a fair comparison of them. We control the size of the training set by randomly choosing 7K sentences (the number of sentences in the DA train) from the en-de NewsTests training set. We train a SIMPLE on this small NewsTests train, and compare its performances to the one of the SIMPLE that was trained on en-de DA. We run each ensemble 3 time with different seeds, to confirm the results. The DA model gets an average correlation of 0.196 with std of 0.024. The COMET model gets an average correlation of 0.219 and std 0.005. Running the COMET model on the full data results with correlation of 0.652. These results suggest that although the COMET score seems to be easier to predict, the size of the training dataset is still the most important factor.

	#sents	in domain	out of domain
NewsTests	28,887	0.64	0.30
bible-uedin	48,705	0.72	0.36
GlobalVoices	55,822	0.55	0.35
Tatoeba	197,381	0.36	0.25

Table 5: The correlations vary between the datasets, both for the in-domain and out-of-domain. Pearson's r of the predictions of the models with/without training on the dataset.

C.2 Out-Of-Domain Datasets

As discussed in §5, our datasets differ on their domain. We take instances of SIMPLE, with no ensembles this time (to reduce training time), and train them on ChrF++ augmentations. First, to measure in-domain, we train and test one instance on each one of the dataset. To measure out-of-domain, for each dataset we train another instances, this time avoiding this dataset during training, and only testing on it (e.g. train on NewsTests, bible-uedin and GlobalVoices, test on Tatoeba). We use the development sets for the testing.

We present the results in Table 5. The correlations vary between the datasets, both for the in-domain and for the out-of-domain. We conclude that similar to other NLP tasks, the domain plays an important role.

C.3 Learning from the Reference

All of our experiments were focused on the ability of our PreQuEL model to give predictions on the expected quality of the translated sentence given the source sentence only. Can it do the same given the reference sentence instead?

In cases where we have a good quality reference (e.g., parallel corpus) we assume the source and reference sentences hold the same content, so it makes sense to expect the answer to this question to be yes. However, it is possible that the challenge of translating from one language to another is more directly related to some features on the source side.

To test that without the artifact of the input language, we use de-en COMET data (See §7.2). This way we will have our references in English.

First Dataset	
1.sub-obj	0.845
2.obj-subject	0.799
Second Dataset	
1.sub-obj	0.865
2.obj-sub	0.870
3.re-sub-obj	0.869
4.re-obj-sub	0.851

Table 6: Mean score predictions for the syntax datasets.

sents pair	corr
1.sub-obj with 2.obj-sub	**0.901**
1.sub-obj with 3.rev-sub-obj	0.687
1.sub-obj with 4.rev-obj-sub	0.680
2.obj-sub with 3.rev-sub-obj	0.563
2.obj-sub with 4.rev-obj-sub	0.641
3.rev-sub-obj with 4.rev-obj-sub	**0.936**

Table 7: Correlation between all sentences versions, for the second dataset.

The results for one de-en SIMPLE COMET instance that was trained on references is 0.639, similar to a model that was trained on the source.

C.4 Outperforming Existing 'PreQuEL' Model

We train and test SIMPLE Aug on the dataset that Sun et al. used, the QE dataset from WMT2019. This dataset uses HTER to score the quality. They reported a correlation of 0.400, while we mange with intertraining on COMET to achieve correlation of 0.422.

The QE dataset from WMT2019 containes 13,442 training samples, much more than the 7k of the WMT2020. Therefore, although we did improve the results over theirs, the augmentation gain is smaller than we showed in §6 for the WMT2020 DA.

D Word Ordering full results

Tables 6 presents the mean scores for both datasets. Table 7 present the correlation between all sentence versions of the second dataset.

E NL-Augmenter transformations

GenderSwap This transformation swaps all gendered words in a given sentence with their counterparts. Names are also randomly swapped. For example "Bob wants to become a programmer, as his father" is transformed to "Alice wants to become a programmer, as her mother".

TenseTranformation This transformation converts sentences from one tense to the other, for example, "My father goes to gym every day" is transformed to "My father went to gym every day".

RandomDeletion This transformation randomly remove each word of a sentence or paragraph with a probability p.

YesNoQuestionPerturbation This perturbation turns English statements into yes-or-no questions. For example, "He also begins an affair with Veronica Harrington, who bails him out." is transformed to "Does he also begin an affair with Veronica Harrington, who bails him out? Yes."

ChangePersonNamedEntities This transformation acts like a perturbation which changes the name of the person. For example, from "John" to "Cathy".

MultilingualBackTranslation This transformation translates a given sentence from a given language into a pivot language and then back to the original language.

ReplaceNumericalValues This transformation looks for numerical values in the text and replaces it with another random value of the same cardinality. For example, "6.9" may be replaced by "4.2", or "333" by "789".

YodaTransformation. This transformation modifies sentences to flip the clauses such that it like "Yoda Speak". For example, "You still have much to learn" is transformed to "Much to learn, you still have".

F Determining Whether to MT - Threshold 90

Figure 3 presents the Precision/Recall curve for both en-de and et-en development datasets, with threshold 90.

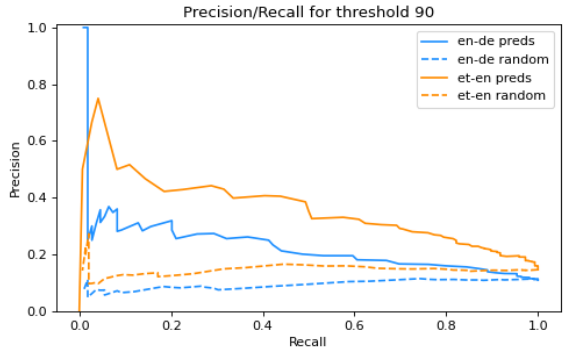

Figure 3: Precision/Recall curve for en-de and et-en development datasets, for the prediction of whether a DA threshold of over 90 would be given to the translation of the input sentence. The blue plots are comparable and so are the orange (but not cross-color comparisons).