Glass-Box Quality Estimation for Neural Machine Translation

João Pinto Correia de Moura
joaopcmoura@tecnico.ulisboa.pt

Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
October, 2021

Abstract. Quality Estimation has become increasingly relevant in the last few years for practical and confidence-aware Machine Translation applications, with recent advancements in the field of Natural Language Processing having enabled new approaches to the task. Despite the great improvements that state-of-the-art Quality Estimation systems boast, most overlook a promising source of information: the translation system under evaluation is treated as a black box, with only its input and output being regarded. In this thesis, we introduce a method which allows for the integration of information extracted from the internal mechanisms of Machine Translation models, into the training process of Quality Estimation models, which we call Glass-Box Quality Estimation. First, in order to extract this internal information, we leverage existing model uncertainty quantification methods based on Monte Carlo dropout, which recent work has shown to yield features highly relevant to estimating the quality of machine translated text. We then propose a novel model architecture based on the Predictor-Estimator framework, and an accompanying method to integrate the extracted features into the model's training procedure. Finally, we provide an empirical evaluation based on six language pairs in the context of the *WMT Quality Estimation Shared Task*, with encouraging results. Our analysis of the proposed model suggests various directions for future improvements.

Keywords: Deep Learning, Natural Language Processing, Quality Estimation, Uncertainty Quantification

1 Introduction

The field of Machine Translation (MT) has undergone a big transformation in the past years, both as a research field and industry. The replacement of Statistical MT by Neural MT (NMT) is widespread in commercial settings, and translation engines are now almost ubiquitously powered by Deep Neural Networks. The improvement of translation quality obtained from the use of these models, more user-friendly tools and higher demand for translation has made common the use of MT models in the translation industry; this setting brings added importance to the development of solutions for a different set of problems, for example:

- which segments need revision by a human translator?
- how much effort will be needed to fix a poorly translated segment?
- which one of many translations made by different models should be picked as the best one?

The Quality estimation (QE) task seeks to address these questions; it consists of predicting the quality of a system's output for a given input, without any information or reference about the expected output, therefore being aimed at MT models in use (Specia et al., 2010). This type of system can be designed for prediction at different granularity levels, from word or sentence, to paragraph or document.

The field of QE, like many others in Natural Language Processing, has benefited greatly from advancements such as the Transformer architecture (Vaswani et al., 2017) and its offspring (BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a), for example), which made it so that models capable of generating powerful contextualized word and sentence embeddings, pre-trained in

many different languages, are generally available for research purposes. These models are typically refered to as pre-trained Language Models in the literature. Transfer learning methods enable the employment of these models' strong Natural Language Understanding capabilities on downstream tasks.

In parallel with the above, unsupervised QE saw interesting advancements hosted by the revisiting of *glass-box* QE features - commonly used in the field's early days (Blatz et al., 2004; Quirk, 2004; Gamon et al., 2005). Leveraging uncertainty quantification methods, Gal and Ghahramani (2016) and Fomicheva et al. (2020) showed that features engineered from the mechanisms internal to state-of-the-art Transformers (in the Machine Translation setting), constitute a rich source of information on translation quality, competitive with supervised QE methods. These newly proposed features are much different to those previously extracted from their Statistical MT ancestors, both because of the process used to extract and engineer them, but also due to the neural structures that originate them, much deeper and denser in information.

This works focuses on combining the advancements explained above, and proposes a method to integrate internal information from NMT systems into the training of state-of-the-art QE models. In the process, we participated in the *Quality Estimation Shared Task* - very much related to our efforts - and published a paper (Moura et al., 2020) on the *2020 Conference on Machine Translation (WMT)*, using the models and method we developed on part of our submission, and obtaining leaderboard results.

2 Quality Estimation

2.1 Word and Sentence-Level Tasks

The two QE sub-tasks we will be exploring are those of predicting quality labels/scores for words and sentences. Different techniques are applied at a document-level, which are not included in the scope of this work.

2.1.1 Word-Level

Word-level QE focuses on predicting quality labels - *OK* or *BAD* - for all tokens in a translated sentence. If we consider that no internal information was leveraged about the NMT system that generated the translation - a black-box approach -, the task can be described as learning to predict the right label (or class) $\boldsymbol{c}$, given a sequence of words in a source language $\boldsymbol{X} = x_1, x_2, ..., x_M$, and its machine translation in a target language $\boldsymbol{Y} = y_1, y_2, ..., y_T$, or $p(\boldsymbol{c}|\boldsymbol{X}, \boldsymbol{Y})$. Moreover, the source sentence is also frequently labelled and used for training in the exact same way, effectively teaching the model to also evaluate the source sentence for correctness; both objectives are usually trained in conjunction.

2.1.2 Sentence-Level

Sentence-level QE, on the other hand, is perfectly described as a regression task: meaning, the objective is to predict a real value, that quantifies the quality of a sentence's machine translation. Making the same consideration of a black-box approach as we did for word-level QE, this means predicting a value $\hat{\boldsymbol{y}} = f(\boldsymbol{X}, \boldsymbol{Y})$. The function f will be implemented by our system.

A few different measurements have been used in the literature as a true label for quality. The most common one historically is the Human-Targeted Translation Error Rate (HTER) (Snover et al., 2006); this indicator is defined as:

$$\text{HTER} = \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions} + \text{Shifts}}{\text{Number of words in reference}} \quad (1)$$

Each inserted/deleted/modified word or punctuation mark counts as one error, and shifting a string of any number of words, by any distance, also counts as one error. The reference translation, in this case, is the post-edit created by a human translator. A post-edit is nothing more than a "fix", or correction, to a machine translated piece of text.

Another commonly used quality indicator is the Direct Assessment score. This is a score directly obtained from a professional translator's assessment of a particular translated sentence, and is normally defined as a score of 0-100, 100 being a perfect translation.

2.2 Predictor-Estimator Architecture

The Predictor-Estimator, originally proposed by Kim et al. (2017a), is an RNN-based architecture very inspired upon the Encoder-Decoder (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Bahdanau et al., 2016; Sutskever et al., 2014), which effectively standardized the end-to-end neural model approach still used in modern QE.

The concept of this architecture is a two-step training process, each focused on a component of the whole system: first, the Predictor is pre-trained

using parallel data, i.e source sentences and reference translations. This component is nothing more than a word prediction model; very similar to the Encoder-Decoder, the big difference to it is that target context is used on both sides of the word that is being predicted, instead of just the preceding context. In the initial pre-task of training the Predictor, for each sample a random target word is replaced with X, and the model then tries to predict the original target word by conditioning on the whole source and target context. The authors assumed such a task to enable the word prediction model to transfer useful knowledge for QE (later found to be true with transformer-based "word-predictors" like BERT, and holding up for many other language tasks), which is passed forward in the form of *Quality Estimation Feature Vectors* (QEFV's). The second step is to train the Estimator, this time with QE data (source sentences, machine translations and quality annotations). This component takes the Predictor's output, and is responsible for estimating the quality of the word/sentence in question.

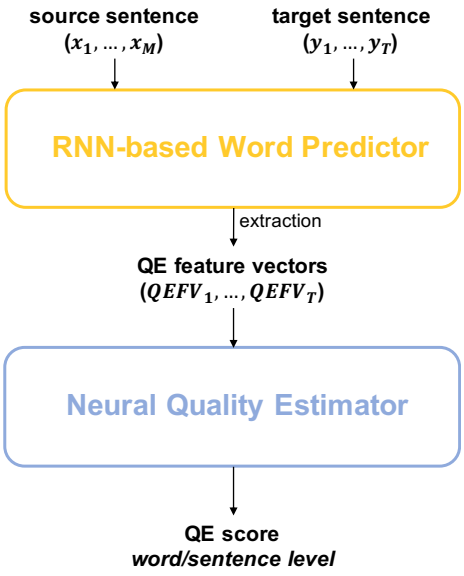

Figure 1: Simplified schematic of the Predictor-Estimator architecture; the Predictor is pre-trained on parallel data (generally more available), and then the whole system - Predictor + Estimator - is trained on QE data.

3 Implemented Systems

We implemented our proposed model architecture (both with and without the integration of glass-box features) on top of the open-source OpenKiwi QE framework[1].

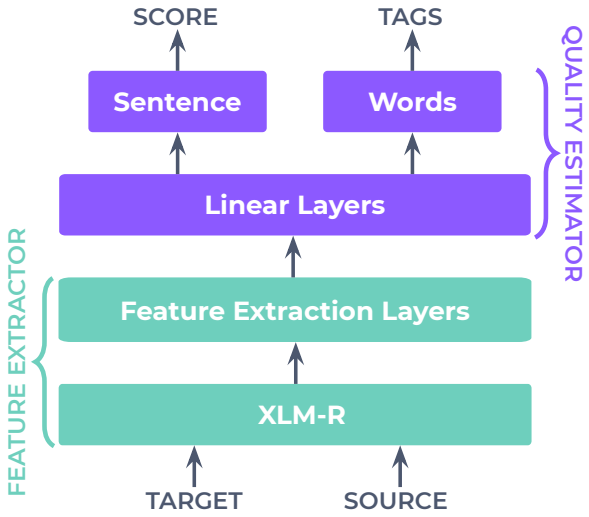

Figure 2: General architecture of the implemented OpenKiwi-based systems.

3.1 Base Kiwi System

Given the success in doing transfer learning with pre-trained Language Models observed in the previous edition of the *WMT Quality Estimation Shared Task* (Kepler et al., 2019), we decided to use XLM-Roberta (Conneau et al., 2020a) models as the Predictor component in our architecture, either `base` (∼270Mparameters with 12-layers, 768-hidden-state, 3072 feed-forward hidden-state, 8 attention heads), or `large` versions (∼550M parameters with 24-layers, 1024-hidden-state, 4096 feed-forward hidden-state, 16 attention heads). We chose XLM-Roberta (called XLM-R from here on), due to its reported state-of-the-art performance on downstream cross-lingual tasks and based on preliminary experiments. Also, since XLM-R is trained on 100 languages (including the ones comprising the dataset that was used), this allowed us to optimize on one system for all language-pairs.

The architecture follows the overall pattern introduced originally in the Predictor-Estimator model, comprising a "Feature Extractor" module with a "Quality Estimator" module on top. Figure 2 depicts this general architecture.

The Feature Extractor module consists of a pre-trained XLM-R model and feature extraction methods on top, such that features for the target sentence, the target tokens, and the source tokens are returned separately. Source and target sentences are passed as inputs in the format `<s> target`

[1] OpenKiwi

</s> <s> source </s>. Output features for tokens in the target sentence are averaged and then concatenated with the classifier token embedding (first <s> in the input), and returned as sentence features.[2]

For the Quality Estimator module we used linear layers instead of a bi-LSTM (as used by Kim et al. (2017b)), since initial experiments showed similar performance. Additional linear layers were stacked on top for each output type: target words, target gaps, source words, and sentence regression.

For the plain OpenKiwi experiments (i.e. using a black-box approach) we used the XLM-R base model and a Quality Estimator block with three feed-forward layers. Hyper-parameter search[3] was performed for each language pair and both subtasks in the *Shared Task*, which will be detailed in a further section. These systems will be referred to as OPENKIWI-BASE through the rest of the paper.

3.2 Glass-Box QE

3.2.1 Glass-Box Features

Recent work on MT confidence estimation (Fomicheva et al., 2020) showed that useful information coming from an MT system, obtained as a by-product of translation, can be competitive with supervised black-box QE models in terms of correlation to human judgements of translation quality, in settings where the labeled data is scarce. The approach described in Fomicheva et al. (2020) requires access to the MT system that produced the translations (unlike the black-box regime). The *2020 WMT Quality Estimation Shared Task* was, in this context, a fitting opportunity alongside which to develop our work. Not only were novel datasets made available that related to our objective, but most importantly the models which created the translations in those datasets were as well.

In our work, we investigated how to combine the richness of this extra information coming from the provided Neural MT (NMT) system with the strength of state-of-the-art approaches to supervised QE. To this end, we extract features (referred to as *glass-box features* henceforth) using the output probability distribution obtained from (i) a standard deterministic NMT and (ii) using uncertainty quantification.

For (ii) we use Monte Carlo Dropout (Gal and Ghahramani, 2015) as a way of circumventing the miscalibration problem of Deep Neural Networks (Guo et al., 2017) and obtaining measures indicative of the model's uncertainty. This method consists of applying dropout at test time before every layer in the network, performing several forward passes for the same inputs (each affected differently by the applied dropout), and collecting posterior probabilities generated by the model. Features are then created from these probabilities, which are used to represent model uncertainty.

We obtain 7 different features for each sentence in each language pair's dataset (see Section 4.1), the first 3 via (i) and the last 4 via (ii) (full details are in Fomicheva et al. (2020)):

- TP - sentence average of word translation probability
- Softmax-Ent - sentence average of softmax output distribution entropy
- Sent-Std - sentence standard deviation of word probabilities
- D-TP - average TP across $N (N = 30)$ stochastic forward-passes
- D-Var - variance of TP across N stochastic forward-passes
- D-Combo - combination of D-TP and D-Var defined by $1 - \text{D-TP}/\text{D-Var}$
- D-Lex-Sim - lexical similarity - measured by METEOR score (Banerjee and Lavie, 2005) - of MT output generated in different stochastic passes.

Feature extraction was implemented as an extension of the Fairseq[4] open-source sequence modelling toolkit.

3.2.2 QE Model and Feature Integration

Different configurations were attempted in order to introduce the extracted glass-box features into the OpenKiwi system. The best empirical performance was observed with a simple method: we reduced the dimension of the pooled sentence features output from XLM-R by about five fold (onto

[2] Even though XLM-R was not trained on the Next Sentence Prediction objective (therefore not using the classification token in its original pretraining), preliminary experiments showed that concatenating inputs, average pooling, and using the classification token resulted in better performance compared to feeding source and target separately and extracting sentence features with other strategies (only pooled target, only the classifier token, classifier token + pooled source, and others).

[3] Hyper-parameters that were searched are: learning rate, dropout, number of warmup steps, and number of freeze steps.

[4] https://github.com/pytorch/fairseq

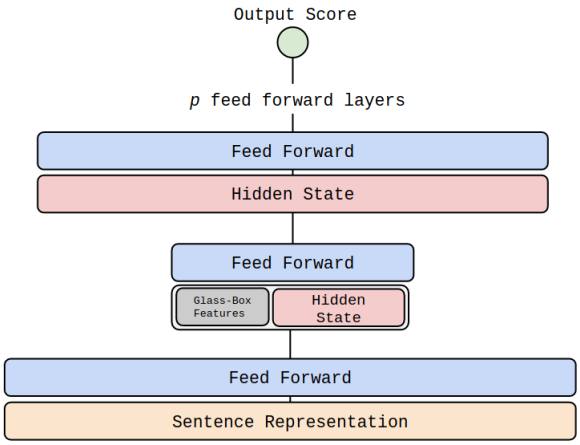

Figure 3: Architecture of the "Quality Estimator" module modified to include *glass-box features*.

`bottleneck_size`), creating a dimensional bottleneck and forcing a more compact sentence representation, and then concatenated the seven extracted glass-box features to this hidden state, followed by an expansion back to a higher dimensional state of `hidden_size`. The result is used as input feature for regression on the sentence score, employing `p` progressively smaller feed-forward layers (halving in size). A visualization of this process can be seen in Figure 3.

The glass-box features were individually normalized a priori, according to their mean and variance in the training dataset, allowing for their integration in the network's training in a scale-independent way.

Although glass-box features weren't extracted or used on a word level, both the sentence and word Estimators will share the weights of the feed-forward layer that succeeds the Predictor in a multitask setting - that is, when the model is trained on both tasks simultaneously, with the word and sentence-level losses being summed as a global loss before updating the network's weights. We posit that the word Estimator can benefit from the influence that glass-box features have on the sentence Estimator, by means of backpropagation and the updating of this feed-forward layer's weights.

For this final version of our system, XLM-R `large` was used instead of `base` version. From here on ou we will call KIWI-GLASS-BOX to the system as described here, but for our final comparison we will also refer to KIWI-LARGE as the same system, but without using the glass-box features.

4 Experiments and Results

4.1 Tasks, Dataset and Model Resources

The dataset we used for developing the models presented in this thesis had a strong influence on the direction of the work itself. A part of the 2020 edition of the *WMT Quality Estimation Shared Task* (Specia et al., 2020), there were two tasks relevant to our work: task 1, for predicting Direct Assessment scores (sentence-level), and task 2, for predicting Post-Editing effort (OK/BAD word labels, and sentence-level HTER 2.1.2).

Both tasks share the same dataset, newly sourced mainly from Wikipedia articles. It includes six language-pairs - 2 high, 2 medium, and 2 low-resource - namely English-German and English-Chinese (high), Romanian-English and Estonian-English (medium), Nepalese-English and Sinhala-English (low). An extra high-resource language-pair was added, Russian-English, however separated from the rest when it comes to content, being comprised of Russian Reddit forums (75%), and Russian WikiQuotes (25%). Datasets for all language-pairs were divided into 7K sentences for training, and 1K sentences for development. Only a subset of the full dataset was annotated and made available for task 2, specifically for the English-German and English-Chinese language pairs.

For task 1, each sentence was annotated following the FLORES setup (Guzman et al., 2019), which presents a form of DA, created with the purpose of standardizing scores, so that the evaluator's rating distribution is taken into account and does not form bias in a labelled dataset. At least 3 professional translators rated all sentences from 0-100, and their scores were standardized using the z-score by rater, defined as $z = \frac{x-\mu}{\sigma}$, where x is a raw DA score, and σ is the standard deviation of ratings for a given evaluator. The scores were then averaged for each translation, the final sentence-level quality label which we will refer to as z-mean score henceforth.

Furthermore, as previously mentioned the NMT models used to create the dataset were made available, so that system-internal information could be exploited in this task. These are standard Transformer models, with 6 encoder blocks and 6 decoder blocks.

All *Shared Task* submissions of KIWI-GLASS-BOX to task 1 were created by simple linear ensembles, combining 5 of the models obtained through hyper-parameter search for each language pair. We

	Feature	Language Pair						
		En-De	En-Zh	Ro-En	Et-En	Ne-En	Si-En	Ru-En
(i)	TP	0.0993	0.2808	0.5951	0.3992	0.3653	0.3658	0.3658
	Softmax-Ent	0.0858	0.2919	0.5595	0.3546	0.4133	0.4077	0.3790
	Sent-Std	0.0691	**0.3252**	0.5049	0.3985	0.3669	0.3912	0.3510
(ii)	D-TP	**0.1078**	0.3158	**0.6404**	**0.4936**	0.3905	0.3797	**0.4441**
	D-Var	0.0782	0.1943	0.3550	0.2780	0.2336	0.2338	0.2329
	D-Combo	0.0487	0.1259	0.2620	0.1335	0.2938	0.2244	0.2013
	D-Lex-Sim	0.0994	0.2903	0.6210	0.3940	**0.4751**	**0.4318**	0.4092

Table 1: Pearson correlation (r) between the employed *glass-box features* and human DA's for every language pair in task 1 (validation set) - best results are in bold.

used the validation set predictions of these 5 models to train a LASSO regression model. However, since we do not possess labels for the test set, these ensembles were trained using k-fold cross-validation ($k = 10$) on the validation set.

In task 2, we trained our model in a multitask setting (as described in Section 3.2.2), training on all three subtasks at the same time: target tags, source tags (using the embeddings that correspond to words in the source sentence), and sentence score. These three outputs are then predicted by the model in a single run. The best model was selected by the highest sum of the resulting three metrics on the validation set.

4.2 Evaluation Metrics

For both sentence-level tasks (DA and HTER prediction), the standard evaluation metric for the task was used, namely the Pearson's correlation coefficient. This is measure of linear correlation between two sets of data, defined as $\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$, where X and Y are the z-mean scores and model predictions, and cov is the covariance between them, and σ each one's standard deviation.

For the word-level task, again the standard evaluation metric was used, which is the Matthews correlation coefficient (MCC), defined by:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \quad (2)$$

where TP, TN, FP and FN are true positives, true negatives, false positives and false negatives, in order. While there is no perfect way of describing a confusion matrix by a single number, MCC is generally regarded as one of the best at doing so.

4.3 Results

4.3.1 Glass-Box Features

First, we confirm the original premise (Fomicheva et al., 2020) that the extracted glass-box features are on par with supervised quality estimation methods, in terms of correlation with human judgement. To this end, we extract the features from the provided models (Section 4.1), using the validation sets for each language pair. We then calculate the Pearson r correlation for each feature-language pair in task 1; results can be seen in Table 1.

As expected, features obtained using uncertainty quantification–based (feature group (ii)) consistently display higher correlations across all language-pairs, `D-TP` being the most effective for high and medium resource languages, and `D-Lex-Sim` for low resource languages. This is in accordance with intuition, given that MT models trained on low-resource languages have had less data points to train and converge on, and might therefore create more variable outputs for the same source, when affected by dropout.

We will refer to the best correlation achieved by any glass-box feature for each language pair as BEST GB FEATURE in the following section.

4.3.2 Glass-Box QE

Most comparisons we draw from the obtained results in the following paragraphs are expressed for task 1; this task was initially worked on more deeply, and was the one used as experimentation and feedback mechanism to understand the impact of the developed method. Once validated, the method was applied in a straightforward way to task 2.

Results for both tasks are shown in Tables 2 and 3; analyzing them, we can answer a series of questions that help assess the different components and experiments we led:

Are glass-box features good unsupervised quality estimators?

As we alluded to in section 4.3.1, the extracted features by themselves achieve a very comparable - and for some language pairs, even better - perfor-

Pair	System	Target MCC		Source MCC		Pearson	
		Val	Test	Val	Test	Val	Test
En-De	KIWI-GLASS-BOX	**0.460**	**0.465**	**0.357**	**0.349**	**0.618**	**0.633**
	OPENKIWI-BASE	0.445	0.432	0.330	0.324	0.561	0.531
	(*)OpenKiwi 1.0	-	0.358	-	0.266	-	0.392
En-Zh	KIWI-GLASS-BOX	0.567	0.567	**0.348**	0.287	**0.691**	**0.651**
	OPENKIWI-BASE	**0.576**	**0.575**	0.298	0.287	0.615	0.593
	(*)OpenKiwi 1.0	-	0.509	-	0.270	-	0.506

Table 2: Task 2 word and sentence-level results on the validation and test sets. Results for OPENKIWI-BASE and KIWI-GLASS-BOX were obtained from a single model trained by multitasking on the 3 different subtasks. (*) Baseline results on the validation set were not made available by the *Shared Task* organizers.

Pair	System	Pearson	
		VAL	TEST
En-De	(*)**KIWI-GLASS-BOX-ENSEMBLE**	0.5715	0.5230
	KIWI-GLASS-BOX	0.5263	-
	KIWI-LARGE	0.4794	-
	OPENKIWI-BASE	0.3499	0.2670
	BEST GB FEATURE	0.1078	-
	Openkiwi 1.0	-	0.1455
En-Zh	(*)**KIWI-GLASS-BOX-ENSEMBLE**	0.5711	0.4940
	KIWI-GLASS-BOX	0.5461	-
	KIWI-LARGE	0.5258	-
	OPENKIWI-BASE	0.4199	0.3460
	BEST GB FEATURE	0.3252	-
	OpenKiwi 1.0	-	0.1902
Ro-En	(*)**KIWI-GLASS-BOX-ENSEMBLE**	0.8968	0.8910
	KIWI-GLASS-BOX	0.8841	-
	KIWI-LARGE	0.8790	-
	OPENKIWI-BASE	0.6672	0.7080
	BEST GB FEATURE	0.6404	-
	OpenKiwi 1.0	-	0.6845
Et-En	(*)**KIWI-GLASS-BOX-ENSEMBLE**	0.7697	0.7700
	KIWI-GLASS-BOX	0.7611	-
	KIWI-LARGE	0.7496	-
	OPENKIWI-BASE	0.6728	0.6900
	BEST GB FEATURE	0.4936	-
	OpenKiwi 1.0	-	0.4770
Ne-En	(*)**KIWI-GLASS-BOX-ENSEMBLE**	0.7994	0.7920
	KIWI-GLASS-BOX	0.7804	-
	KIWI-LARGE	0.7711	-
	OPENKIWI-BASE	0.6987	0.6040
	BEST GB FEATURE	0.4751	-
	OpenKiwi 1.0	-	0.3860
Si-En	(*)**KIWI-GLASS-BOX-ENSEMBLE**	0.6896	0.6390
	KIWI-GLASS-BOX	0.6604	-
	KIWI-LARGE	0.6521	-
	OPENKIWI-BASE	0.5727	0.5650
	BEST GB FEATURE	0.4318	-
	OpenKiwi 1.0	-	0.3737
Ru-En	(*)**KIWI-GLASS-BOX-ENSEMBLE**	0.7391	0.7670
	KIWI-GLASS-BOX	0.7137	-
	KIWI-LARGE	0.6938	-
	OPENKIWI-BASE	-	-
	BEST GB FEATURE	0.4441	-
	OpenKiwi 1.0	-	0.5479

Table 3: Task 1 results on the validation and test sets for all language pairs in terms of Pearson's r correlation. Systems in **bold** were officially submitted. (*) Lines with an asterisk use LASSO regression to tune ensemble weights on the validation set, therefore their numbers cannot be directly compared to the other models.

mance than the best approach from previous years' *Shared Task* submissions, OpenKiwi 1.0. This is all the more of an impressive result, considering that these features are extracted in a completely unsupervised manner, and points to potential use cases where labelled data is not available - either at all, or in a big enough quantity to train an accurate QE model.

Are pre-trained contextualized embeddings a better choice than training a Predictor from scratch?

Using pre-trained contextualized embeddings with XLMRoberta proves to greatly outperform the baseline system OpenKiwi 1.0 in both tasks, even when not taking XLM-R large into account. The fact that RNN's are replaced by a Transformer architecture, the optimization for multilingual pre-training implemented in XLM-R (Conneau et al., 2020a), and the sheer amount of data that XLM-R is pre-trained with, all contribute to this difference. It also highlights the virtues of using transfer-learning in NLP tasks; pre-training a neural model with the size and amount of data that has been proven to be required for powerful language representation is impossible in most cases, and taking advantage of the computational expense incurred by large research entities enables explorations such as the one developed for this thesis.

What effect does increasing the Predictor's capacity have?

Switching from using XLM-R Base to Large as the Predictor component has a very strong impact in performance, with the increase in correlation averaging 11,3% across language-pairs (tested on task 1 only). This is in line with findings from the original XLM-R paper (Conneau et al., 2020b), which indicated that adding capacity to a multilingual model alleviates the *curse of multilinguality* (degrading performance caused by training on many

languages), and results in higher performance for the same number of languages involved in the training process.

Do glass-box features positively inform the training of QE models?

The developed Kiwi-Glass-Box approach consistently increases performance across language pairs and in both tasks. Sentence-level improvements are more visible in predicting HTER (task 2), with the increase averaging 6,6% for both language-pairs, but DA prediction performance benefits nonetheless (task 1), averaging 2% across language-pairs, and as high as 4.7% in the case of En-De. The glass-box features are leveraged by the model during training, resulting in a stronger correlation with human judgement than either one separately. It is curious to note that, even though features extracted for high-resource language pairs show the lowest correlations independently, the QE models trained on those language pairs make the best use of them, judging by performance increase. This could be a factor of the language itself, or alternatively, the translation patterns of a more extensively trained NMT model, learned by the QE model, which might find in uncertainty measurements more signal for determining translation quality.

In the multi-task setting of task 2, results show that is possible for word-level performance to be positively influenced by added information at the sentence-level. Apart from the case of target label prediction for En-Zh, the word-level task and its corresponding component in the model architecture (the final binary classification block) benefits from the extra information on which the shared layers are trained (refer to Section 3.2.2).

All in all, the proposed method improves performance across the board on both word and sentence-level prediction.

5 Conclusions

This thesis was developed with the objective of enhancing state-of-the-art QE models, by allowing them to leverage internal features from NMT models - or glass-box features.

Our starting point was the set of glass-box features introduced in Fomicheva et al. (2020), originally used directly as unsupervised quality estimators, and proven to be effective at representing NMT model uncertainty. We began by implementing the extraction of these features for the NMT models used .

Then, in Section 3.1 we proposed a QE model architecture on which to apply these features, leveraging the Natural Language Understanding capabilities of a multi-lingual, pre-trained Language Model (XLM-Roberta). Independently, this architecture yielded results that surpassed the performance of RNN-based QE systems - the state-of-the-art up until this point -, as we show in Section 4.3.2.

Finally, in Section 3.2.2 we developed a method to introduce glass-box features into the proposed model's training process. We validated our design choices, confirm the independent relevance of the features as quality indicators, and at last show that the QE model's performance in the sentence-level task consistently increases across language pairs, when using them as extra input information. We also show that, although features are extracted on a sentence-level granularity only, multi-task learning paired with weight sharing between sentence and word-level Estimator components has a positive influence on the model's performance in predicting word-level labels. These conclusions are part of the evaluation drawn from results across two tasks and six language pairs, obtained as the result of our participation in the *WMT Quality Estimation Shared Task* (Moura et al. (2020)), and according to the proposed metrics.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. 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.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computation Linguistics*.

Yarin Gal and Zoubin Ghahramani. 2015. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of The 33rd International Conference on Machine Learning*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-level MT evaluation without reference translations: beyond language modeling. In *Proceedings of the 10th EAMT Conference: Practical applications of machine translation*, Budapest, Hungary. European Association for Machine Translation.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *ArXiv*, abs/1706.04599.

Francisco Guzman, 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. pages 6100–6113.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M. Amin Farajian, António V. Lopes, and André F. T. Martins. 2019. Unbabel's participation in the WMT19 translation quality estimation shared task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 78–84, Florence, Italy. Association for Computational Linguistics.

H. Kim, H. Jung, Hong-Seok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017a. Predictor-estimator. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17:1 – 22.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017b. Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation. In *Conference on Machine Translation (WMT)*.

João Moura, Miguel Vera, Daan van Stigt, Fabio Kepler, and André F. T. Martins. 2020. IST-unbabel participation in the WMT20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1029–1036, Online. Association for Computational Linguistics.

Christopher B. Quirk. 2004. Training a sentence-level machine translation confidence measure. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Matthew Snover, Bonnie J. Dorr, R. Schwartz, and L. Micciulla. 2006. A study of translation edit rate with targeted human annotation.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André FT Martins. 2020. Findings of the wmt 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.