Combining Quality Estimation and Automatic Post-editing to Enhance Machine Translation Output

Rajen Chatterjee chatterjee@fbk.eu
Matteo Negri negri@fbk.eu
Marco Turchi turchi@fbk.eu
Fondazione Bruno Kessler, Trento, Italy

Frédéric Blain f.blain@sheffield.ac.uk
Lucia Specia l.specia@sheffield.ac.uk
University of Sheffield, Sheffield, UK

Abstract

We investigate different strategies for combining quality estimation (QE) and automatic post-editing (APE) to improve the output of machine translation (MT) systems. The joint contribution of the two technologies is analyzed in different settings, in which QE serves as either: *i)* an *activator* of APE corrections, or *ii)* a *guidance* to APE corrections, or *iii)* a *selector* of the final output to be returned to the user. In the first case (QE as activator), sentence-level predictions on the raw MT output quality are used to trigger its automatic correction when the estimated (TER) scores are below a certain threshold. In the second case (QE as guidance), word-level binary quality predictions ("good"/"bad") are used to inform APE about problematic words in the MT output that should be corrected. In the last case (QE as selector), both sentence- and word-level quality predictions are used to identify the most accurate translation between the original MT output and its post-edited version. For the sake of comparison, the underlying APE technologies explored in our evaluation are both phrase-based and neural. Experiments are carried out on the English-German data used for the QE/APE shared tasks organized within the First Conference on Machine Translation (WMT 2016). Our evaluation shows positive but mixed results, with higher performance observed when word-level QE is used as a selector for neural APE applied to the output of a phrase-based MT system. Overall, our findings motivate further investigation on QE technologies. By reducing the gap between the performance of current solutions and "oracle" results, QE could significantly add to competitive APE technologies.

1 Introduction

In recent years, the steady progress of machine translation (MT) technology motivated research on a number of ancillary tasks. MT's wide adoption, especially in the translation industry, has in fact raised new challenges, which are not only related to model training, optimization, adaptation and evaluation, but also to aspects that are external to the core translation approach. Among them, quality estimation (QE) and automatic post-editing (APE) deal respectively with the possibility of *predicting* the quality of MT output and *correcting* it via downstream post-processing. QE (Specia et al., 2010) is motivated by the need for estimating output quality at

run-time, that is when reference-based evaluation is unfeasible (the typical scenario when MT is deployed in production). APE (Simard et al., 2007) is motivated by the need of improving MT systems' output in black-box conditions, in which the translation models are not accessible for internal modification, retraining or adaptation (a typical situation for companies that rely on third-party MT systems).

Both QE and APE have been successfully explored as standalone tasks in previous work, in particular within the well-established framework of the Conference on Machine Translation (WMT[1]). In six editions of the WMT QE shared task (2012-2017), the MT quality prediction problem has been formulated in different ways (*e.g.* ranking, scoring) and attacked at different levels of granularity (sentence/phrase/word-level). Constant state-of-the-art advancements are making QE a more appealing technology, for instance to enhance the productivity of human translators operating with computer-assisted translation tools (Bentivogli et al., 2016). Three rounds of the APE shared task at WMT (2015-2017) followed a similar trend, with improvements that reflect a steady progress of the underlying technology developed by participants.

Despite the growing interest in the two tasks and the fact that the proposed evaluation exercises shared the same training/test data, previous research on both topics has mainly followed independent paths. As a consequence, the potential usefulness of leveraging the two technologies to achieve better MT has been scarcely explored and no systematic analysis of the possible combination strategies has been reported. Along this direction, this paper investigates how QE and APE can be jointly deployed to boost the overall quality of the output produced by an MT system, without any intervention on the actual MT technology. By experimenting with the same data, we explore different ways to approach the problem. The main difference between the proposed strategies lies in the degree of integration between the two technologies. A light integration is pursued when QE is used either to trigger the automatic correction of machine-translated text (QE as *activator*, see Section 3.1) or to validate an automatic correction by comparing it with the original MT output (QE as *selector*, see Section 3.3). A tighter integration is pursued when QE is used to inform the automatic correction process by identifying problematic passages in the machine-translated text (QE as *guidance*, see Section 3.2). Depending on the applied strategy, QE predictions can be done at different levels of granularity. In this first exploration we focus on the two most studied levels, namely sentence and word levels, leaving for future work the exploitation of phrase-level estimates (Logacheva and Specia, 2015; Bojar et al., 2016).

Overall, this paper presents the following main contributions:

- The first systematic analysis of different strategies for the integration of QE and APE towards better MT quality;
- Experimental results, computed on the same public test set, indicating that state-of-the-art QE methods can improve APE and that, in turn, their joint contribution can boost MT output quality;
- A verification of our findings, based on "oracle" quality scores, indicating a large room for improvement conditioned to the reliability of QE predictions. This motivates further research on the task.

2 Background

This section introduces the two MT ancillary tasks relevant to this paper (QE and APE), and summarizes related research. It also gives an overview of the few initiatives in which the interaction between the QE and APE technologies has been previously explored.

[1] http://www.statmt.org/wmt17/

2.1 Quality Estimation

Machine translation quality estimation is the task of predicting the quality of machine-translated text at run-time, without relying on hand-crafted reference translations (Specia et al., 2010). Its possible uses include: *i)* deciding whether a given translation (or portions of it) is good enough for publishing as-is or needs post-editing by professional translators (*e.g.* in a computer-assisted translation environment), *ii)* informing readers of the target language about the reliability of a translation, *iii)* selecting the best translation among options from multiple MT and/or translation memory systems.

QE is usually cast as a supervised learning task, in which systems trained on (*input*, *output*, *quality_label*) triplets have to predict a quality label for unseen (*input*, *output*) test instances. In previous works, this problem has been attacked from different perspectives that, in recent years, reflected the formulation of various sub-tasks proposed within the WMT shared evaluation framework.[2] Major differences concern the granularity and type of the predictions, as well as the underlying learning paradigm. The granularity of QE predictions can range from the document to the sentence, phrase or word level. Relevant to this paper are the sentence and word levels, the former being also the most widely studied type of QE.

In sentence-level QE, the required predictions can be post-editing effort estimates (*e.g.* the expected number of editing operations required to correct the MT output), post-editing time estimates (*e.g.* the time required to make an automatic translation acceptable), ranking of multiple translation options, binary ("good"/"bad") scores or Likert-scale scores (*e.g.* 1-to-5 scores indicating overall translation quality as perceived by a human). Depending on the type of prediction required, the proposed supervised learning approaches span from classification to regression and ranking. Together with the batch learning solutions that characterize the majority of the proposed approaches, recent work also explored the application of online and multitask learning methods (Turchi et al., 2014; C. de Souza et al., 2015) targeting flexibility and robustness to domain differences between training and test data.

In word-level QE, the required predictions can be either binary "good"/"bad" labels for each MT output token, or more fine-grained multi-class labels indicating the type of error occurring in a specific word. The problem, initially approached as a sequence labelling task (Luong et al., 2013), has then been successfully tackled with neural solutions that now represent the state-of-the-art (C. de Souza et al., 2014; Kreutzer et al., 2015; Martins et al., 2016; Kim et al., 2017).

For the experiments discussed in this paper (see Section 4.2) sentence-level and word-level quality estimates are obtained with the winning systems at the WMT 2016 QE shared task, described respectively in (Kozlova et al., 2016) and (Martins et al., 2016). For a comprehensive overview of the evolution of the QE task and the proposed approaches, we refer the reader to the rich WMT literature (Callison-Burch et al., 2012; Bojar et al., 2013, 2014, 2015, 2016, 2017).

2.2 Automatic Post-editing

Automatic post-editing is the task of correcting errors produced by a machine translation system, typically by exploiting knowledge acquired from human post-edits. APE research dates back to the work of (Knight and Chander, 1994; Simard et al., 2007), in which the problem is approached as a "monolingual translation" task. Similar to MT, in which systems are trained on bilingual data consisting of parallel (*source*, *MT*) pairs, APE systems are built by learning from monolingual data consisting of "parallel" (*source*, *MT*, *post_edited MT*) triplets, in which a source sentence and its raw MT output are associated with a human-corrected version of the translation. Under this general formulation, previous work concentrated on several aspects, with outcomes that indicate a steady evolution of the approaches.

[2] http://www.statmt.org/wmt17/quality-estimation-task.html

In terms of methods, early works rely on a statistical phrase-based approach (Simard et al., 2007), possibly integrating source information to reduce systems' tendency to over-correct and enhance their precision (Béchara et al., 2011; Chatterjee et al., 2015). More recently, large performance gains comparable to those achieved in MT by the adoption of neural approaches have also been observed in APE (Junczys-Dowmunt and Grundkiewicz, 2016), especially with multi-source neural systems (Chatterjee et al., 2017a) enhancing decoding with source language information.

In order to evaluate the effects of combining QE with different APE technologies, the experiments discussed in Section 4.3 are performed with both phrase-based (Chatterjee et al., 2016) and neural-based (Junczys-Dowmunt and Grundkiewicz, 2016) architectures. The latter is the winning system at the WMT 2016 APE shared task.[3] For a comprehensive overview of the evolution of the APE task and the proposed approaches, we refer the reader to the aforementioned WMT literature.

2.3 Combination of QE and APE

So far, the interaction and the possible joint contribution of QE and APE technology has been scarcely explored. This is particularly surprising if we consider that both the tasks can rely on the type of same training data consisting of (*source*, *MT*, *post_edited MT*) parallel triplets, which in principle allow for knowledge transfer and model sharing.

Building on this consideration, (Martins et al., 2017) exploited the synergies between the two related tasks by using the output of an APE system as an extra feature to boost the performance of a neural QE architecture. The intuition is that word-level quality labels can be automatically obtained through TER (Snover et al., 2006) alignments between the translated and the post-edited sentence (used as "pseudo human post-edit"). The resulting APE-based QE system achieves state-of-the-art performance at the word and sentence-level QE tasks on the WMT 2015 and WMT 2016 data sets.

Another line of research that is closer to our work has focused on improving APE performance by leveraging word-level QE predictions. In (Hokamp, 2017), this is done by incorporating word-level features as factors in the input (*i.e.* a concatenation of different word embedding representations) to a neural APE system. By taking a different approach, Chatterjee et al. (2017b) explore a "guided-decoding" mechanism to guide a neural APE system with word-level binary QE judgments. The idea of constraining the decoding process (*i.e.* forcing the system to keep "good" tokens in the APE output) is the basis for the integration approach described in Section 3.2: "QE as APE guidance".

A third solution for the integration of QE and APE is explored in (Chatterjee et al., 2016), in which sentence-level quality predictions are used to select the best translation between the raw output or the correction produced by a factored phrase-based APE model. This integration strategy is the basis for the integration approach described in Section 3.3: "QE as MT/APE selector".

3 QE and APE Integration Strategies

QE and APE can be combined in different ways to enhance MT quality. In the following sections we identify and evaluate the following three alternative strategies:

1. QE as APE *activator*: QE predictions are used to trigger APE decoding when the estimated quality of the MT segment is below a certain threshold;

2. QE as *guidance* for the APE decoder: QE labels are used to inform the APE decoding process by discriminating which tokens in the MT output should be kept or changed;

[3] http://www.statmt.org/wmt16/ape-task.html

3. QE as *selector*: QE predictions are used to identify the best alternative between the raw MT and its automatically corrected version. This decision can be performed either at the level of entire sentences or for portions of the two alternative outputs.

3.1 QE as APE activator

In this first scenario, QE is used to take a decision on whether activating APE or not. This is done by running a sentence-level QE on the MT segment to predict its TER score, and then setting a threshold on this prediction. If the predicted TER is below the threshold,[4] the translation will be considered as good enough and the application of an automatic post-editing step unnecessary. Instead, if the predicted TER is above the threshold, the APE decoder is run, and its output is shown to the end user.

3.2 QE as APE guidance

In the "QE as APE activator" strategy, the APE decoder is not directly aided by the QE information to improve its performance, because QE is applied before running the correction process. This prevents QE from supporting APE in addressing specific problems such as over-correction, a well known issue of APE systems: they tend to "re-translate" the raw MT output, even when it is already good. Indeed, as evidenced by the last editions of the APE shared task at WMT (Bojar et al., 2016), the submitted systems performed unnecessary corrections that could either be penalized by automatic evaluation against references consisting of minimal human corrections (*i.e.* only those that are strictly necessary) or, in the worst case, even worsen the MT segment.

To overcome this limitation, an alternative strategy is a QE+APE pipeline in which fine-grained word-level QE judgments for each MT token are passed as additional information to the APE decoder, with the aim of guiding it to change only the words marked as "bad" (or, in other terms, to keep the correct MT tokens in the APE output). While implementing this "QE as guidance" strategy under the phrase-based framework is straightforward, its application to neural APE decoding requires some adaptation work.

In the case of phrase-based APE decoding, the XML markup technique can be easily applied. With this approach, word-level QE labels are directly passed to the APE decoder by specifying a fixed translation for a specific span of the source sentence. If the predicted label is "good", the suggested span contains the original MT words (*i.e.* the decoder is forced to preserve them in the final output). If the label is "bad", instead, the corresponding MT word is not marked, thus giving the decoder the freedom to modify it. Among the different strategies to combine the suggested translation options and those proposed by the APE model, in Section 5 we experiment with the *inclusive* setting, in which the proposed options compete against the other translation candidates in the phrase table.

When the APE system is based on a neural decoder, the XML markup strategy is implemented following the guided decoding mechanism proposed in (Chatterjee et al., 2017b). Differently from a standard neural decoder that predicts one word at a time by leveraging the previously predicted word, the context and the previous hidden states, the guided decoder is enhanced by *i)* a method to prioritize the suggested word in the beam search, *ii)* a look-ahead mechanism to avoid duplicates of the suggested words, and *iii)* a strategy to generate continuous and discontinuous target phrases. More details about the algorithm are available in (Chatterjee et al., 2017b). Similarly to the XML markup, the "good" labels are transformed in suggested spans containing the MT words, thus pushing the decoder towards using them. For "bad" word-level predictions, in contrast, the decoder does not receive any constraint and is free to produce the most probable output.

[4] Recall that TER is an error metric, so lower scores indicate better translations.

3.3 QE as MT/APE selector

A third possible strategy is to exploit QE predictions at a downstream level, after APE processing. After the APE decoder has generated its output, QE can be used to determine the best output option between the original MT and the post-edited segments. In our experiments, this is done in two ways, namely: *i)* by annotating only the MT segment with word-level QE labels, or *ii)* by annotating both the MT and the APE outputs at sentence or word-level.

In the first scenario, the MT and APE segments are first word aligned. Then, by using the MT output as a backbone, words are retained or modified based on their QE predicted labels. MT words labelled as "good" will be retained, while those marked as "bad" will be replaced by the corresponding alignments from the APE output.

In the second scenario, both the MT and APE sentences are labelled with sentence-level QE predictions (TER scores), and the one with the lower predicted TER score is selected as final output. To make the decision process more robust, a threshold τ can be set on the difference in TER between the two segments. For instance, if the goal is to take a conservative approach favouring the MT system, such threshold can be set such that APE outputs are selected only if their predicted TER is much lower than the MT (TER_{MT} - $\text{TER}_{APE} > \tau$).

When both the MT and APE sentences are annotated with word-level binary labels, the tokens marked as "good" are selected from one of the two segments. In detail, the MT and APE segments are first word aligned, then the MT is taken as backbone. For each MT word, the QE labels are compared. If the MT label is "good" and the APE is "bad", the MT word is taken. If the MT label is "bad" and the APE is "good", the APE word is selected. In case the annotations are both either "bad" or "good", the MT word is chosen. In the case where either MT or APE word is aligned to NULL with a "good" label, the word is added to the final output. Although this technique is rather simple, as shown in Section 5, it results in competitive performance.

4 Experimental Settings

4.1 Data

The experiments in this paper are performed using the English-German (En-De) datasets released for the APE and QE shared tasks at WMT 2016 (Bojar et al., 2016), wich are a subset of a larger collection presented in (Specia et al., 2017). Each item consists of a triplet (*source*, *MT*, *post_edited MT*) obtained by first translating the source sentence with a phrase-based statistical MT (PBSMT) system, and then by post-editing the translated segment. The data belongs to the Information Technology domain and the post-edits are created by professional translators. The training set contains 12,000 triplets, the development set, 1,000 and the test set 2,000 items. For the word-level QE task, the three sets have 21.4%, 19.54% and 19.31% "bad" labels, showing an unbalanced distribution towards the "good" quality tokens.

4.2 QE systems

For generating the sentence-level predictions, the best system submitted at the 2016 QE shared task is used (Kozlova et al., 2016). It consists in a pipeline of two regressors, where the first one, given a set of features, predicts the BLEU score (Papineni et al., 2002) and the second one, given the predicted BLEU value, predicts the TER score. Several features are combined, including features extracted from the parse trees of the sentences, pseudo-references, back-translation, web-scale language model, and word alignments.

At word-level, the best performing system at the 2016 QE shared is used (Martins et al., 2016). It is a stacked architecture that combines three neural models: one feed-forward and two recurrent ones. The predictions of these three models are added as features in a feature-based linear sequential model. Syntactic dependency-based features are combined with the baseline features released by the task organizers.

In our experiments we test different QE-APE integration strategies using either the predicted QE annotations produced by the aforementioned systems or the gold (ORACLE) labels released by the task organizers. The main idea behind the use of oracle labels is to evaluate the improvements in MT quality that would be possible given a perfect QE predictor.[5]

4.3 APE systems

The outputs of two APE systems are used in the "QE as APE activator" and "QE as MT/APE selector" experiments. These systems are the best PBSMT APE system (Chatterjee et al., 2016), and the best neural APE approach (Junczys-Dowmunt and Grundkiewicz, 2016) at the second edition of the APE shared task.

The fist method relies on a factored PBSMT decoder that combines the classic monolingual approach (*MT* $\rightarrow$ *pe* translation) with the context-aware method (*MT#src* $\rightarrow$ *pe*) that improves the decoding process with information from the source sentence. This system is enhanced with the addition of several language models including part-of-speech-tag, neural class-based and statistical word-based. In addition to these components, the *primary* submitted run includes a quality estimator as MT/APE selector, which did not result in better performance compared to a system without it. To avoid any bias in our evaluation, we use the submission that does not include the QE component (*i.e.* the submitted *contrastive* run). It relies on the Moses platform (Koehn et al., 2007), a 5-gram word-based statistical language model, and 8-gram POS-tag and class-based language models. obtained with statistical and neural language model toolkits.

The neural system is an attentional encoder-decoder model trained with sub-word units. The primary submission is an ensemble of monolingual (*MT* $\rightarrow$ *pe*) and cross-lingual (*source* $\rightarrow$ *pe*) systems combined in a log-linear model. A task-specific feature based on string matching is added to the log-linear combination to control the faithfulness of the APE results with regards to the input. Differently from the PBSMT system, the neural model requires a large quantity of training data. This data is obtained by a "round-trip translation" process that generates source and MT segments starting from the reference sentences. In total, $\sim$4 million artificial triplets are used to train a generic neural APE system that is then fine-tuned on the task-specific (APE) data.

In the "QE as APE activator" and "QE as MT/APE selector" strategies, the interaction between the QE and the APE systems is performed before and after decoding. For this reason, we directly take the submissions of the two systems to the APE shared task. In the tighter integration explored in the "QE as APE guidance" experiments, the test set needs to be post-edited using the QE labels. For this purpose, the PBSMT APE system, that is an instance of the Moses toolkit with standard parameters, is trained only on the task-specific data. This is different from the best PBSMT APE system at WMT 2016, because we only use a statistical word-level language model. For the guided decoder, the implementation and settings from Chatterjee et al. (2017b) are used. This means that the neural APE model has been trained by the authors of the winning system at the WMT 2016 APE shared task (Junczys-Dowmunt and Grundkiewicz, 2016) using the large "round-trip translation" dataset, and then adapted to task-specific data. The network parameters are: word embedding dimensionality of 600, hidden unit size of 1,024, maximum sentence length of 50, batch size of 80, and vocabulary size of 40K. The network parameters are optimized with Adadelta (Zeiler, 2012).

For the guided decoder, the best value step of the look-ahead mechanism is defined on the development set. The data is segmented using the Byte-Pair Encoding (BPE) technique (Sennrich et al., 2016). Each QE word-level annotation is projected to all the subword units. We only use a single *MT* $\rightarrow$ *pe* model instead of an ensemble of models.

[5]It is important to note that there are several perfectly valid translations of the same input text, so the gold QE predictions that we use are a subset of possible oracle labels generated based on the available reference sentence.

It is important to note that both the APE systems are strong. In fact, they significantly improve over the MT output (respectively +2.64 and +5.54 BLEU points for the phrase-based and the neural-based system) and, compared to the best system at WMT 2017 that uses twice the size of the task-specific data and leverages the multi-source neural architecture, the performance gap is limited to 2 BLEU score points.

5 Results and Discussion

In this section, the light and tight integration of QE and APE are evaluated to identify conditions where translation quality can be enhanced. In all the experiments, the combined QE and APE systems are compared against *i)* the original MT baseline and *ii)* the APE system without QE (official WMT submissions when possible, our implementations in the results reported in Table 2). Both PBSMT and neural APE are considered. In the APE shared task at WMT 2016, both systems outperform significantly the MT baseline. The performance of all the systems is evaluated in terms of BLEU score against the reference translation.

QE as APE activator. In this set of experiments we investigate if the light integration of QE, as a way to trigger the automatic correction of machine translated texts, is able to improve translation quality. For this purpose, both sentence-level QE predictions and ORACLE values are computed for each MT sentence. If these values are larger than a threshold, then the APE decoder is run to improve the translation. Different TER thresholds on the QE scores are tested on the development set (*i.e.* ranging from 0 to 100 with step 5) and the best performing value is applied to the test set. The performances of the MT, APE (WMT systems), APE+QE predictions and APE+QE$_{ORACLE}$ are reported in Table 1.

	APE System	
	PBSMT	Neural
MT	62.11	62.11
APE (WMT systems)	64.75	67.65
APE + QE	64.47	67.19
APE + QE$_{ORACLE}$	64.58	67.56

Table 1: QE as APE activator (BLEU scores). These results are obtained using a threshold of 10 TER points.

As it can be seen in Table 1, for the two APE+QE configurations, a marginal drop in performance indicates that using sentence-level QE does not help this task. The use of the ORACLE labels (last row in table) is marginally better than the use of predicted QE scores, but it is still worse than using the top performing APE systems without QE. This is true for both the PBSMT and neural APE systems. Our intuition is that sentence-level QE scores provide information that is too coarse-grained, which does not give any hint to the APE system about what is wrong in the MT output and how difficult it is to correct it. For instance, not running the APE decoder when the TER (either oracle or predicted) is small does not mean that APE could not correct the (few) errors present in the MT segment.

"QE as APE guidance". To better support APE, a tighter integration of the two technologies is obtained by injecting word-level QE annotations directly into the decoder. This is done by using the XML markup in the case of the PBSMT model and by means of guided decoding for the neural model. Results are reported in Table 2.[6]

[6]The APE results are different compared to the ones reported in Table 1 because our PBSMT APE only uses the word-level language model, and our neural APE is a single *MT* $\rightarrow$ *pe* system instead of an ensemble.

	APE System	
	PBSMT	Neural
MT	62.11	62.11
APE (our implementation)	63.47	65.25
APE + QE	63.57	65.50
APE + QE$_{ORACLE}$	63.78	67.03

Table 2: QE as APE guidance (BLEU scores).

Differently from the sentence-level QE predictions, the word-level predictions are effective and their use results in a small but significant gain in performance over the APE system alone. This improvement also exists with the neural decoder, which is already a stronger APE module on its own. When using the ORACLE labels, there are further improvements in BLEU score. This is more evident for the neural system (+1.53 BLEU), bringing it closer to the ensemble APE system (67.75 BLEU) shown in Table 1. A possible explanation of this larger gain compared to the PBSMT APE is that the higher generalization capability of the neural approach, which forces the APE system to perform a large number of changes, can be kept under control using information from QE. Moreover, the guided decoder proposes a tighter integration of the QE annotations than the XML markup, which is not able to decode phrases spanning across the suggested words and those for which a modification is required.

It is worth noticing that the observed relative improvements are obtained on top of simpler implementations of PBSMT and neural APE systems. For this reason, they are not directly comparable with the APE results reported in Tables 1, 3, 4 and 5, which are obtained from participants' official submissions at WMT 2016. However, while lower than those achieved by the top WMT 2016 systems, these positive results suggest that it would be worth testing the QE-guided APE approach on more competitive state-of-the-art APE solutions.

"QE as MT/APE selector". In the last round of experiments we use QE information after the APE decoding. For this configuration, two solutions are explored. In the first one (see Table 3), only the MT segment is labelled with word-level QE annotations (predicted and ORACLE). In the second one, both the MT and APE sentences are annotated with sentence or word-level QE information (Tables 4 and 5 respectively). The results in Table 3 show that both predicted and ORACLE sentence-level annotations of the MT output can enhance the quality obtained by the standard APE (WMT systems). Similarly to the "QE as APE guidance" approach, the neural APE is more sensitive to QE information, achieving a significant +0.2 (predictions) and +1.56 (ORACLE) BLEU scores over the standard APE. We hypothesize that these results stem from the tendency of neural APE systems to perform a large number of modifications on the MT output which are not always correct. QE information on the MT thus limits unnecessary changes made by the APE module.

	APE System	
	PBSMT	Neural
MT	62.11	62.11
APE (WMT systems)	64.75	67.65
APE + QE	64.87	67.86
APE + QE$_{ORACLE}$	65.13	69.21

Table 3: QE as MT/APE selector (BLEU scores). Word-level QE annotations are produced only for the MT segment.

When both the MT and the APE segments are annotated with the sentence-level QE scores, a threshold is set to decide whether to show the MT or the APE translation to the end user. The best results, reported in Table 4, are obtained with τ equal to 5 and 1, respectively for the predicted and ORACLE TER values. These experiments confirm that the use of the predicted sentence information is not useful: both the PBSMT and neural APE+QE systems produce outputs that are worse than the standard APE (WMT systems). when using the ORACLE annotations, the BLEU scores achieved are better than those for the APE system on its own, and in line with the performance obtained by ORACLE word-level QE information on the MT segments (last row in Table 3). This indicates that better QE scores would be helpul in this setting.

	APE System	
	PBSMT	Neural
MT	62.11	62.11
APE (WMT systems)	64.75	67.65
APE + QE	64.49	66.49
APE + QE$_{ORACLE}$	65.26	69.50

Table 4: QE as MT/APE selector (BLEU scores). Sentence-level QE annotations both on the MT and APE segments.

The final experiment consists in annotating both the MT and the APE segments with word-level QE information and defining a simple strategy to merge the two outputs (see Section 3.3). Table 5 shows that both PBSMT and neural APE systems take advantage of the QE labels, slightly improving over the APE systems on their own. Unsurprisingly, larger gains with both techniques are obtained using the ORACLE annotations: a performance boost of +0.68 and +3.34 BLEU scores respectively. Again, the neural APE achieves the best performance and largest potential improvement, confirming that the large variability of the applied changes is indeed an advantage, and that it can be kept under control using information from QE.

	APE System	
	PBSMT	Neural
MT	62.11	62.11
APE (WMT systems)	64.75	67.65
APE + QE	64.83	67.79
APE + QE$_{ORACLE}$	65.51	71.13

Table 5: QE as MT/APE selector (BLEU scores). Word-level QE annotations both on the MT and APE segments.

Overall, by taking into consideration all the experiments we report in this paper, our main findings can be summarized as follows:

- Integration strategies exploiting word-level QE seem to be more promising than those based on sentence-level QE. Our results show that sentence-level QE information is too coarse to support APE decoding, while having the QE annotations on each MT and/or APE token can help to enhance overall translation quality.

- At word-level, predicted QE labels yield limited but constant gains, up to $\sim$0.2 BLEU points over standard APE systems. These values are small, but support the intuition that QE and APE integration has some potential.

- ORACLE results indicate that there is large scope for improvement if better QE systems can be designed. Although it is not guaranteed that a QE model could achieve the ORACLE performance, it is interesting to notice that increasing the quality of the QE annotations results in significant translation quality improvements.

6 Conclusion

We proposed a systematic analysis of different techniques to combine QE and APE to achieve better MT quality. These strategies range from light integration, where QE is used either to trigger automatic post-editing or to compare the APE with the original MT segment, to tighter integration, in which QE annotations are directly used to guide the inner workings of the APE decoder. Our experiments show that QE can help APE to produce better MT outputs. Among the proposed strategies, "QE as guidance" and "QE as selector" lead to improvements in MT quality. The use of neural APE and word-level QE, on both MT and APE, results in the largest gains over the top WMT 2016 APE system (+0.2 BLEU score with the predicted QE annotations, and +3.34 BLEU score with the ORACLE labels).

These findings motivate further research in this area, which can be explored with different directions. First, all the proposed strategies have been applied to the output of a PBSMT MT system, but the recent advancements in neural MT call for testing our integrated approaches also on the output of a NMT model. Second, the progress of deep learning methods in APE and QE suggests that the development of a single end-to-end model that would simultaneously leverage both technologies could be beneficial (*e.g.* pointer network (Vinyals et al., 2015) for neural APE).

References

Béchara, H., Ma, Y., and van Genabith, J. (2011). Statistical Post-Editing for a Statistical MT System. In *Proceedings of the 13th Machine Translation Summit*, pages 308–315, Xiamen, China.

Bentivogli, L., Bertoldi, N., Cettolo, M., Federico, M., Negri, M., and Turchi, M. (2016). On the evaluation of adaptive machine translation for human post-editing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):388–399.

Bojar, O. et al. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Bojar, O. et al. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA.

Bojar, O. et al. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal.

Bojar, O. et al. (2016). Findings of the 2016 Conference on Machine Translation. In *Proceedings of the 1st Conference on Machine Translation*, pages 131–198, Berlin, Germany.

Bojar, O. et al. (2017). Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the 2nd Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark.

C. de Souza, J. G., González-Rubio, J., Buck, C., Turchi, M., and Negri, M. (2014). FBK-UPV-UEdin Participation in the WMT14 Quality Estimation shared-task. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, Baltimore, MD, USA.

C. de Souza, J. G., Negri, M., Ricci, E., and Turchi, M. (2015). Online Multitask Learning for Machine Translation Quality Estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 219–228, Beijing, China.

Callison-Burch, C. et al. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.

Chatterjee, R., C. de Souza, J. G., Negri, M., and Turchi, M. (2016). The FBK Participation in the WMT 2016 Automatic Post-editing Shared Task. In *Proceedings of the 1st Conference on Machine Translation*, pages 745–750, Berlin, Germany.

Chatterjee, R., Farajian, M. A., Negri, M., Turchi, M., Srivastava, A., and Pal, S. (2017a). Multi-source Neural Automatic Post-Editing: FBK's Participation in the WMT 2017 APE Shared Task. In *Proceedings of the 2nd Conference on Machine Translation*, pages 630–638, Copenhagen, Denmark.

Chatterjee, R., Negri, M., Turchi, M., Federico, M., Specia, L., and Blain, F. (2017b). Guiding Neural Machine Translation Decoding with External Knowledge. In *Proceedings of the 2nd Conference on Machine Translation*, pages 157–168, Copenhagen, Denmark.

Chatterjee, R., Weller, M., Negri, M., and Turchi, M. (2015). Exploring the Planet of the APEs: a Comparative Study of State-of-the-art Methods for MT Automatic Post-Editing. In *Proceedings of ACL '15*, pages 156–161, Beijing, China.

Hokamp, C. (2017). Ensembling factored neural machine translation models for automatic post-editing and quality estimation.

Junczys-Dowmunt, M. and Grundkiewicz, R. (2016). Log-linear Combinations of Monolingual and Bilingual Neural Machine Translation Models for Automatic Post-Editing. In *Proceedings of the 1st Conference on Machine Translation*, Berlin, Germany.

Kim, H., Lee, J.-H., and Na, S.-H. (2017). Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the 2nd Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark.

Knight, K. and Chander, I. (1994). Automated postediting of documents. In *Proceedings of the 12th National Conference on Artificial Intelligence*, pages 779–784, Seattle, WA.

Koehn, P. et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings ACL '07*, Stroudsburg, PA, USA.

Kozlova, A., Shmatova, M., and Frolov, A. (2016). Ysda participation in the wmt'16 quality estimation shared task. In *Proceedings of the 1st Conference on Machine Translation*, pages 793–799, Berlin, Germany.

Kreutzer, J., Schamoni, S., and Riezler, S. (2015). Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, pages 316–322, Lisbon, Portugal.

Logacheva, V. and Specia, L. (2015). Phrase-level quality estimation for machine translation. In *Proceedings of the 12th International Workshop on Spoken Language Translation, Da Nang, Vietnam*.

Luong, N. Q., Lecouteux, B., and Besacier, L. (2013). LIG System for WMT13 QE Task: Investigating the Usefulness of Features in Word Confidence Estimation for MT. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, pages 386–391, Sofia, Bulgaria.

Martins, A., Astudillo, R., Hokamp, C., and Kepler, F. (2016). Unbabel's participation in the wmt16 word-level translation quality estimation shared task. In *Proceedings of the 1st Conference on Machine Translation*, pages 806–811, Berlin, Germany.

Martins, A., Junczys-Dowmunt, M., Kepler, F., Astudillo, R., Hokamp, C., and Grundkiewicz, R. (2017). Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL '02*, pages 311–318, Philadelphia, Pennsylvania, USA.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of ACL '16*, pages 1715–1725, Berlin, Germany.

Simard, M., Goutte, C., and Isabelle, P. (2007). Statistical Phrase-Based Post-Editing. In *Proceedings of NAACL HLT '07*, pages 508–515, Rochester, New York.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA 2006*, pages 223–231, Cambridge, Massachusetts, USA.

Specia, L., Harris, K., Blain, F., Burchardt, A., Macketanz, V., Negri, M., and Turchi, M. (2017). Translation quality and productivity: A study on rich morphology languages. In *Proceedings of the 16th Machine Translation Summit*, Nagoya, Japan.

Specia, L., Raj, D., and Turchi, M. (2010). Machine Translation Evaluation versus Quality Estimation. *Machine translation*, 24(1):39–50.

Turchi, M., Anastasopoulos, A., C. de Souza, J. G., and Negri, M. (2014). Adaptive Quality Estimation for Machine Translation. In *Proceedings of ACL '14*, pages 710–720, Baltimore, Maryland.

Vinyals, O., Fortunato, M., and Jaitly, N. (2015). Pointer networks. In *Advances in Neural Information Processing Systems 28*, pages 2692–2700.

Zeiler, M. D. (2012). ADADELTA: An Adaptive Learning Rate Method. *CoRR*, abs/1212.5701.