Findings of the WMT 2018 Shared Task on Quality Estimation

Lucia Specia and **Frédéric Blain**
Department of Computer Science
University of Sheffield, UK
{l.specia,f.blain}@sheffield.ac.uk

Varvara Logacheva
Neural Networks and Deep Learning Lab
MIPT, Moscow, Russia
logacheva.vk@mipt.ru

Ramón F. Astudillo
L2F, INESC-ID-Lisboa
Lisbon, Portugal
ramon@astudillo.com

André Martins
Unbabel & Instituto de Telecomunicações
Lisbon, Portugal
andre.martins@unbabel.com

Abstract

We report the results of the WMT18 shared task on Quality Estimation, i.e. the task of predicting the quality of the output of machine translation systems at various granularity levels: word, phrase, sentence and document. This year we include four language pairs, three text domains, and translations produced by both statistical and neural machine translation systems. Participating teams from ten institutions submitted a variety of systems to different task variants and language pairs.

1 Introduction

This shared task builds on its previous six editions to further examine automatic methods for estimating the quality of machine translation (MT) output at run-time, without the use of reference translations. It includes the (sub)tasks of word-level, phrase-level, sentence-level and document-level estimation. In addition to advancing the state of the art at all prediction levels, our goals include:

- To study the performance of quality estimation approaches on the output of neural MT systems. We do so by providing datasets for two language pairs where source segments were translated by both statistical phrase-based and neural MT systems.
- To study the predictability of missing words in the MT output. To do so, for the first time we provide data annotated for such errors at training time.
- To study the predictability of source words that lead to errors in the MT output. To do so, for the first time we provide source segments annotated for such errors at the word level.
- To study the effectiveness of manually assigned labels for phrases. For that we provide a dataset where each phrase was annotated by human translators.
- To investigate the utility of detailed information logged during post-editing. We do so by providing post-editing time, keystrokes, as well as post-editor ID.
- To study quality prediction for documents from errors annotated at word-level with added severity judgements. This is done using a new corpus manually annotated with a fine-grained error taxonomy, from which document-level scores are derived.

This year's shared task provides new training and test datasets for all tasks, and allows participants to explore any additional data and resources deemed relevant. Tasks make use of large datasets produced either from post-editions or annotations by professional translators, or from direct human annotations. The following text domains are available for different languages and tasks: information technology (IT), life sciences, and product title and descriptions on sports and outdoor activities. In-house statistical and neural MT systems were built to produce translations for the two first domains, while an online system was used for the third domain.

The four tasks are defined as follows: Task 1 aims at predicting post-editing effort at sentence level (Section 5); Task 2 aims at predicting words that need editing, as well as missing words and incorrect source words (Section 6); Task 3 aims at predicting phrases that need editing, as well as missing phrases and incorrect source phrases (Section 7); and Task 4 (Section 8) aims at predicting a score for an entire document as a function of the proportion of incorrect words in such a document, weighted by the severity of the different errors.

Five datasets and language pairs are used for different tasks (Section 4): English-German (Tasks 1, 2) and English-Czech (Tasks 1, 2) on the IT domain, English-Latvian (Tasks 1, 2) and German-English (Tasks 1, 2, 3), both on the life sciences domain, English-French (Task 4) with product titles and descriptions within the sports and outdoor activities domain.

Participants are provided with a baseline set of features for each task, and a software package to extract these and other quality estimation features, and perform model learning (Section 2). Participants (Section 3) could submit up to two systems for each task and language pair. A discussion on the main goals and findings from this year's task is given in Section 9.

2 Baseline systems

Sentence-level baseline system: For Task 1, QUEST++[1] (Specia et al., 2015) was used to extract 17 MT system-independent features from the source and translation (target) files and parallel corpora:

- Number of tokens in the source and target sentences.
- Average source token length.
- Average number of occurrences of the target word within the target sentence.
- Number of punctuation marks in source and target sentences.
- Language model (LM) probability of source and target sentences based on models built using the source or target sides of the parallel corpus used to train the SMT system.
- Average number of translations per source word in the sentence as given by the IBM model 1 extracted using the SMT parallel corpus, and thresholded such that $P(t|s) > 0.2$ or $P(t|s) > 0.01$.
- Percentage of unigrams, bigrams and trigrams in frequency quartiles 1 (lower frequency words) and 4 (higher frequency words) in the source language extracted from the source side of the SMT parallel corpus.
- Percentage of unigrams in the source sentence seen in the source side of the SMT parallel corps.

These features were used to train a Support Vector Regression (SVR) algorithm using a Radial Basis Function (RBF) kernel within the SCIKIT-LEARN toolkit.[2] The γ, ϵ and C parameters were optimised via grid search with 5-fold cross validation on the training set, resulting in γ=0.01, ϵ = 0.0825, C = 20. This baseline system has been consistently used as the baseline system for all editions of the sentence-level task (Callison-Burch et al., 2012; Bojar et al., 2013, 2014, 2015, 2016, 2017), and has proved strong enough for predicting various forms of post-editing effort across a range of language pairs and text domains for statistical MT systems. This year it is also benchmarked on neural MT outputs.

Word-level baseline system: For Task 2, the baseline features were extracted with the MARMOT tool (Logacheva et al., 2016). These are 28 features that have been deemed the most informative in previous research on word-level QE, mostly inspired by (Luong et al., 2014). This is the same baseline system used in WMT17:

- Word count in the source and target sentences, and source and target token count ratio. Although these features are sentence-level (i.e. their values will be the same for all words in a sentence), the length of a sentence might influence the probability of a word being wrong.
- Target token, its left and right contexts of one word.
- Source word aligned to the target token, its left and right contexts of one word. The alignments were given by the SMT system that produced the automatic translations.
- Boolean dictionary features: target token is a stop word, a punctuation mark, a proper noun, or a number.
- Target language model features:
  - The order of the highest order ngram which starts and end with the target token.
  - The order of the highest order ngram which starts and ends with the source token.
  - The part-of-speech (POS) tags of the target and source tokens.
  - Backoff behaviour of the ngrams (t_{i-2}, t_{i-1}, t_i), (t_{i-1}, t_i, t_{i+1}), (t_i, t_{i+1}, t_{i+2}), where t_i is the target

[1] https://github.com/ghpaetzold/questplusplus

[2] http://scikit-learn.org/

token (backoff behaviour is computed as described by (2011)).

In addition to that, six new features were included which contain combinations of other features, and which proved useful in (Kreutzer et al., 2015; Martins et al., 2016):

- Target word + left context.
- Target word + right context.
- Target word + aligned source word.
- POS of target word + POS of aligned source word.
- Target word + left context + source word.
- Target word + right context + source word.

The baseline system models the task as a sequence prediction problem using the Linear-Chain Conditional Random Fields (CRF) algorithm within the CRFSuite tool.[3] The model was trained using passive-aggressive optimisation algorithm.

We note that this baseline system was only used to predict OK/BAD classes for existing words in the MT output. No baseline system was provided for predicting missing words or erroneous source words.

Phrase-level baseline system: The phrase-level system is identical to the one used in last year's shared task. The phrase-level features were also extracted with MARMOT, but they are different from the word-level features. They are based on the sentence-level features in QUEST++.[4] These are the so-called "black-box" features – features that do not use the internal information from the MT system. The baseline uses the following 72 features:

- Source phrase frequency features:
  - average frequency of ngrams (unigrams, bigrams, trigrams) in different quartiles of frequency (the low and high frequency ngrams) in the source side of the SMT parallel corpus.
  - percentage of distinct source ngrams (unigrams, bigrams, trigrams) seen in the source side of the SMT parallel corpus.
- Translation probability features:
  - average number of translations per source word in the phrase as given by the IBM model 1 extracted using the SMT parallel corpus (with different translation probability thresholds: 0.01, 0.05, 0.1, 0.2, 0.5).
  - average number of translations per source word in the phrase as given by the IBM model 1 extracted using the SMT parallel corpus (with different translation probability thresholds: 0.01, 0.05, 0.1, 0.2, 0.5) weighted by the frequency of each word in the source side of the parallel SMT corpus.
- Punctuation features:
  - difference between numbers of various punctuation marks (periods, commas, colons, semicolons, question and exclamation marks) in the source and the target phrases.
  - difference between numbers of various punctuation marks normalised by the length of the target phrase.
  - percentage of punctuation marks in the target or source phrases.
- Language model features:
  - log probability of the source or target phrases based on models built using the source or target sides of the parallel corpus used to train the SMT system.
  - perplexity of the source and the target phrases using the same models as above.
- Phrase statistics:
  - lengths of the source or target phrases.
  - ratio between the source and target phrase lengths.
  - average length of tokens in source or target phrases.
  - average occurrence of target word within the target phrase.
- Alignment features:
  - number of unaligned target words, using the word alignment provided by the SMT decoder.
  - number of target words aligned to more than one source word.

[3] http://www.chokkan.org/software/crfsuite/

[4] http://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox

  - average number of alignments per word in the target phrase.

- Part-of-speech features:
  - percentage of content words in the source or target phrases.
  - percentage of words of a particular part of speech tag (verb, noun, pronoun) in the source or target phrases.
  - ratio of numbers of words of a particular part of speech (verb, noun, pronoun) between the source and target phrases.
  - percentage of numbers and alphanumeric tokens in the source or target phrases.
  - ratio between the percentage of numbers and alphanumeric tokens in the source and target phrases.

Analogously to the baseline word-level system, we treat phrase-level QE as a sequence labelling task, and model it using CRF from the CRFSuite toolkit and the passive-aggressive optimisation algorithm.

Once more, this baseline system was only used to predict OK/BAD classes for existing phrases in the MT output. No baseline system was provided for predicting missing phrases or erroneous source phrases.

3 Participants

Table 1 lists all participating teams submitting systems to any of the tasks. Each team was allowed up to two submissions for each task variant and language pair. In the descriptions below, participation in specific tasks is denoted by a task identifier (T1 = Task 1, T2 = Task 2, T3 = Task 3, T4 = Task 4).

CMU-LTI (T2):

The CMU-LTI team proposes a Contextual Encoding model for QE. The model consists in three major parts that encode the local and global context information for each target word. The first part uses an embedding layer to represent words and their POS tags in both languages. The second part leverages a one-dimensional convolution layer to integrate local context information for each target word. The third part applies a stack of feed-forward and recurrent neural networks to further encode the global context in the sentence before making the predictions. Syntactic features, such as ngrams, are then integrated to the final feed-forward layer in the neural model. This model achieves competitive results on the English-Czech and English-Latvian word-level QE task.

JU-USAAR (T2):

JU-USAAR presents two approaches to word-level QE: (i) a Bag-of-Words (BoW) model, and (ii) a Paragraph Vector (Doc2Vec) model (Le and Mikolov, 2014). In the BoW model, bag-of-words are prepared from source sentences for each target word appearing in both the MT and PE output in the training data. For every target word appearing in the MT output in the development set, the cosine similarity between the corresponding source sentence and the bag-of-words for the same target word is computed. From this result, a threshold (for the target word) is defined above which the word is retained (i.e., considered 'OK'). In the Doc2Vec-based approach, for each target word appearing in both MT and PE output in the training data, two document vectors are prepared from (i) the corresponding source sentences and (ii) the bag-of-words (as in the BoW model) of the target word. Next, the similarity between these two document vectors for every target word is computed. From the Doc2Vec similarity score and the corresponding PE decision (i.e., whether or not the target word is retained in the PE in the training dataset), a system level threshold is defined. For the test set sentences, if the Doc2Vec similarity score for a target word exceeds this threshold value, then the target word labelled as 'OK', otherwise it is labelled as 'BAD'.

MEQ (T1)

The Vicomtech team submitted two approaches. uMQE is an unsupervised minimalist approach based on two simple measures of accuracy and fluency, respectively. Accuracy is computed via overlapping lexical translation bags of words, with a set expansion mechanism based on longest common prefixes and surface-defined named entities. Fluency is computed by taking the inverse of cross-entropy, according to an in-domain language model. Both measures are combined

ID	Participating team
CMU-LTI	Carnegie Melon University, US (Hu et al., 2018)
JU-USAAR	Jadavpur University, India & University of Saarland, Germany (Basu et al., 2018)
MQE	Vicomtech, Spain (Etchegoyhen et al., 2018)
QEbrain	Alibaba Group Inc, US (Wang et al., 2018)
RTM	Referential Translation Machines, Turkey (Biçici, 2018)
SHEF	University of Sheffield, UK (Ive et al., 2018b)
TSKQE	University of Hamburg (Duma and Menzel, 2018)
UAlacant	University of Alacant, Spain (Sánchez-Martíínez et al., 2018)
UNQE	Jiangxi Normal University, China
UTartu	University of Tartu, Estonia (Yankovskaya et al., 2018)

Table 1: Participants in the WMT18 Quality Estimation shared task.

via simple arithmetic means on rescaled values, i.e., no machine learning is used. Since it is unsupervised, the method can only be meaningfully evaluated on the ranking task. sMQE uses the same two features as uMQE, but with supervision. A Support Vector Regressor based on these two features is trained on the available data and used to predict QE scores.

QEbrain (T1, T2):

QE brain uses a conditional target language model as a robust feature extractor with a novel bidirectional transformer which is pre-trained on a large parallel corpus filtered to contain "in-domain like" sentences. For QE inference, the feature extraction model can produce not only the high-level joint latent semantic representation between the source and the machine translation, but real-valued measurements of possible erroneous tokens based on the prior knowledge learned from the parallel data. More specifically, it uses the multi-head self-attention mechanism and transformer neural networks (Vaswani et al., 2017) to build the language model. It contains one transformer encoder for the source and a bidirectional transformer encoder for the target. After the feature extraction model is trained, the features are extracted and combined with human-crafted features from the QE baseline system and fed into a Bi-LSTM predictive model for QE. A greedy ensemble selection method is used to decrease the individual model errors and increase model diversity. The bi-LSTM QE model is trained on the official QE data plus artificially generated data and fine-tuned with only the official WMT18 QE data.

RTM (T1, T2, T3, T4):

These submissions build on the previous year's Referential Translation Machine (RTM) approach (Biçici, 2017). RTMs predict data translation between the instances in the training set and the test set using interpretants, data close to the task instances. Interpretants provide context for the prediction task and are used during the derivation of the features measuring the closeness of the test sentences to the training data, the difficulty of translating them, and to identify translation acts between any two data sets for building prediction models. Task-specific quality prediction RTM models are built using the WMT News translation task corpora, taking MT models as a black-box and predicting translation scores independently on the MT model. Multiple machine learning techniques are used and averaged based on their training set performance for label prediction. For sequence classification tasks (T2 and T3), Global Linear Models with dynamic learning (Bicici, 2013) are used.

SHEF (T1, T2, T3, T4):

SHEF submitted two systems per task variant: SHEF-PT and SHEF-bRNN. SHEF-PT is based on a re-implementation of the POSTECH system of (Kim et al., 2017), SHEF-bRNN uses a bidirectional recurrent neural network (bRNN) (Ive et al., 2018a). PT systems are pre-trained using in-domain corpora provided by the organisers. bRNN systems uses two encoders to learn representations of <source, MT> sentence pairs. These representations are used directly to make word-level predictions. A weighted sum over word representations as defined by an attention mechanism is used to make sentence-level predictions. For phrase-level,

a standard attention-based neural MT architecture is used. Different parts of the source sentence are attended to produce MT word vectors. Phrase-level predictions are based on representations computed as the sum of their word vectors. For predicting source tags, the source and MT inputs to the models are swapped. The document-level architecture wraps the sentence-level PT and bRNN architectures. PT systems are pre-trained using either additional in-domain or out-of-domain Europarl data. For the multi-task learning system (SHEF-mtl), weights of sentence-level modules are pre-trained to predict sentence MQM scores.

TSKQE (T1):

The TSKQE submissions represent an extension over the previous UHH-STK submissions to the WMT17 QE shared task, which combine the power of sequence and tree kernels applied on source segments, candidate translation and back-translations of the MT output into the source language. In addition, in order to predict the HTER scores, one of the current submissions also explores pseudo-references, which were obtained by translating the source sentences into the target language using an online MT system. The sequence kernels were applied on the tokenised data, while tree kernels were applied to dependency trees.

UAlacant (T1, T2):

The UAlacant submissions use phrase tables from OPUS[5] and a two hidden layer feed-forward neural network for word-level MT QE. Phrase tables are used to extract features for each word and gap in the machine-translated segment for which quality is estimated. These features are then used together with the baseline features for predicting the need of a deletion or an insertion. The neural network takes as input not only the features for the word and the gap on which a decision is to be made, but also the features of the surrounding gaps and words in a sliding-window fashion within a context window of size three. The predictions made at the word level allow to obtain an approximate HTER score which is used for the submissions to the sentence-level task.

UNQE (T1):

The UNQE submissions employ the unified neural network architecture (UNQE) for sentence-level QE tasks (Li et al., 2018). The approach combines a bidirectional RNN encoder-decoder with attention mechanism sub-network and an RNN into a single large neural network, which extracts the quality vectors of the translation outputs through the bidirectional RNN encoder-decoder, and predicts the HTER value of the translation output by RNN. The input text goes through tokenisation, true casing and sub-word unit segmentation. The models are pre-trained with a large parallel bilingual corpus and fine-tuned with the training data of the sentence-level QE share task. The results submitted are averages of the predicted HTER scores under different dimension settings.

UTartu (T1):

UTartu proposes two methods for the sentence-level task. The first method uses attention weights of a neural MT system applied to each sentence pair to compute the probability of the output sentence under the model (forced-decoding). The confidence of the model is computed via metrics of average entropy of the attention weights per each input/output token. The second method computes the `bleu2vec` metric, which extends BLEU with token or n-gram embeddings, but here the metric is made cross-lingual by means of an unsupervised cross-lingual mapping between the source and target language embedding spaces. Three versions of the resulting metric are used: one based on 3-grams, one with tokens (unigrams) and one with byte-pair encoded sub-words (also unigrams). Both submissions use the 17 standard black-box features implemented in QuEst. QuEst+Attention combines them with the first approach and QuEst+Att+CrEmb3 combines QuEst and both approaches together.

[5] http://opus.nlpl.eu/

4 Datasets

This year we further expand the datasets used in WMT17 by adding: more instances (see Table 2), more languages (four language pairs), more MT architectures (neural and statistical MT), and different types of annotation (manual and extracted from manual post-editing). In addition, new data was collected and provided for Task 4, on a fifth language pair and third text domain.

4.1 Tasks 1 and 2

The initial data was collected as part of the QT21 project[6] and is fully described in (Specia et al., 2017). However, for all language pairs and MT system types, we filtered this data to remove most cases with no edits performed. A skewed distribution towards good quality translations has been shown to be a problem in previous years, and is even more critical with NMT outputs, where up to about half of the MT sentences require no post-editing at all. We kept only a small proportion of HTER=0 sentences in training, development and test sets.

The structure used for the data has been the same since WMT15. Each data instance consists of (i) a source sentence, (ii) its automatic translation into the target language, (iii) the manually post-edited version of the automatic translation, (iv) one or more post-editing effort scores as labels. Professional post-edits are used to extract labels for the two different levels of granularity (word and sentence). Table 2 shows the various resulting datasets for English-German (EN-DE), German-English (DE-EN), English-Latvian (EN-LV) and English-Czech (EN-CS), for both statistical (SMT) and neural (NMT) outputs.

English-German and English-Czech sentences are from the IT domain and were translated by an in-house phrase-based SMT system, and in addition by an in-house encoder-decoder attention-based NMT system for English-German. We note that the original dataset sizes for these languages was 30,000 sentences in total for English-German (per MT system type), and 45,000 for English-Czech. The large reduction in the NMT version of the English-German data indicates the high quality of the NMT system used to produce these sentences: a large number of sentences was filtered out for having undergone no edits by translators.

German-English and English-Latvian sentences are from the life sciences (pharmaceutical) domain and were translated by an in-house phrase-based SMT system, and in addition by an in-house encoder-decoder attention-based NMT system for English-Latvian. The original sentence numbers for these languages were 45,000 and 20,738, respectively (per MT system type).

4.2 Task 3

This task uses a subset of the German-English SMT data from Task 1 (5,921 sentences for training, 1,000 for development and 543 for test) where each phrase (as produced by the SMT decoder) has been annotated (as a phrase) by humans with four labels (see Section 7). This subset was selected after post-editing by filtering out translations with HTER=0 and with a HTER=0.30 and above, and then randomly selecting a subset large enough while fitting the annotation budget. The latter criterion was used to rule out sentences with too many errors, since these are generally too hard or impossible to annotate for errors by humans.

We used BRAT[7] to perform the phrase labelling. The annotator – a professional translator – was given the translations to annotate, along with their respective source sentence. We provided them with a preset environment where all translations were pre-labelled at phrase-level beforehand as OK. The annotator's task was then to change the labels of the incorrect phrases. The labelling was done following a 'pessimistic' approach, where we requested the annotator to only consider a phrase to be OK if all its words were OK. This task has two variants, as we describe later: Task3a, where a phrase annotation is propagate to all of its words and the task is framed as a word-level prediction task; and Task3b, where prediction is done at the phrase level. Table 3 shows the statistics of the resulting datasets for these variants of the task.

Since the data used for this task is a subset of the dataset of that used for Task 1, we selected as test sentences also a subset of the test set for Task 1.

4.3 Task 4

The document-level task data consists of short **product descriptions** translated from English to French, extracted from the Amazon Product Re-

[6] http://www.qt21.eu/

[7] https://brat.nlplab.org/

	Train.		Dev.		Test	
Language pair	# Sentences	# Words	# Sentences	# Words	# Sentences	# Words
DE-EN	25,963	493,010	1,000	18,817	1,254	23,522
EN-DE-SMT	26,273	442,074	1,000	16,565	1,926	32,151
EN-DE-NMT	13,442	234,725	1,000	17,669	1,023	17,649
EN-LV-SMT	11,251	225,347	1,000	20,588	1,315	26,661
EN-LV-NMT	12,936	258,125	1,000	19,791	1,448	28,945
EN-CS	40,254	728,815	1,000	18,315	1,920	34,606

Table 2: Statistics of the datasets used for Tasks 1 and 2: Total number of (source) sentences and words (after tokenisation) for training, development and test for each language pair and MT system type.

Task3a	# Sentences	# Words	# BAD
Train.	5,921	126,508	35,532
Dev.	1,000	28,710	6,153
Test	543	7,464	3,089
Task3b	# Sentences	# Phrases	# BAD
Train.	5,921	50,834	10,451
Dev.	1,000	8,566	1,795
Test	543	4,391	868

Table 3: Statistics of the data used for Task 3. Number of sentences, phrases, words and BAD labels for training, development and test.

	# Documents	# Sentences	# Words
Train.	1,000	6,003	129,099
Dev.	200	1,301	28,071
Test	269	1,652	39,049

Table 4: Statistics of the data used for Task 4. Number of documents, sentences and (target) words for training, development and test.

views dataset (McAuley et al., 2015; He and McAuley, 2016).[8] More specifically, the data is a selection of Sports and Outdoors product titles and descriptions in English which has been machine translated into French using a state of the art online neural MT system. The most popular products (those with more reviews) were chosen. This data poses interesting challenges for machine translation: titles and descriptions are often short and not always a complete sentence. Spans covering one or more tokens were annotated with error labels following fine-grained error taxonomy, as described in more detail in Section 8. The dataset statistics are presented in Table 4. This is the largest ever released collection with word-level errors manually annotated.

[8] http://jmcauley.ucsd.edu/data/amazon/ .

5 Task 1: Predicting sentence-level quality

This task consists in scoring (and ranking) translation sentences according to the proportion of their words that need to be fixed. HTER is used as quality score, i.e. the minimum edit distance between the machine translation and its manually post-edited version.

Labels Three labels were available: percentage of edits need to be fixed (HTER) (primary label), post-editing time in seconds, and counts of various types of keystrokes. The PET tool (Aziz et al., 2012)[9] was used to collect various types of information during post-editing. HTER labels were computed using the TERCOM tool[10] with default settings (tokenised, case insensitive, exact matching only), with scores capped to 1.

Evaluation Evaluation was performed against the true HTER label and/or ranking, using the following metrics:

- Scoring: Pearson's r correlation score (primary metric, official score for ranking system submissions), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
- Ranking: Spearman's ρ rank correlation.

Statistical significance on Pearson r was computed using the William's test.[11]

Results For Task 1, Tables 5, 6, 7 and 8 summarise the results for English–German, German–English, English–Latvian and English-Czech, respectively, ranking participating systems best to worst using Pearson's r correlation as primary key.

[9] https://github.com/ghpaetzold/PET
[10] https://github.com/jhclark/tercom
[11] https://github.com/ygraham/mt-qe-eval

Spearman's ρ correlation scores should be used to rank systems for the ranking variant of the evaluation.

The top two systems for this task, the QEBrain model and UNQE models, show a large performance gap with respect to the rest of the systems, for both SMT and NMT data. It is interesting to note that both systems outperform the SHEF-PT system by a large margin. SHEF-PT is a reimplementation of the POSTECH system, which showed the top performance in 2017.

6 Task 2: Predicting word-level quality

This task evaluates the extent to which we can detect word-level errors in MT output. Often the overall quality of a translated segment is significantly harmed by specific errors in a small number of words. As in previous years, each token of the target sentence is labeled as OK/BAD based on an available post-edited sentence. In addition to this, this year we also took into consideration word omission errors and the detection of words in the source related to target side errors. These types of errors become particularly relevant in the context of NMT systems. The code to produce this new set of tags from any prior WMT corpora is available for download.[12]

Target word labels As in previous years, the binary labels for each target token (OK and BAD) were derived automatically by aligning each machine translated sentence with its post-edited counterpart sentence. The alignment at token-level was performed using the TERCOM tool. Default settings were used and shifts were disabled. Target tokens originating from insertion or substitution errors were labeled as BAD. All other tokens were labeled as OK.

Gap and source word labels To annotate deletion errors, gap 'tokens' between each word and at the beginning of each target sentence were introduced. These gaps tokens were labeled as BAD in the presence of one or more deletion errors and OK otherwise. To annotate the source words related to insertion or substitution errors in the machine translated sentence, the IBM Model 2 alignments from fastalign (Dyer et al., 2013) were used. Each token in the source sentence was aligned to the post-edited sentence. For each token in the post-edited sentence deleted or substituted in the machine translated text, the corresponding aligned source tokens were labeled as BAD. In this way, deletion errors also result in BAD tokens in the source, related to the missing words. All other words were labeled as OK.

Evaluation Analogously to last year's task, the primary evaluation metric is the multiplication of F_1-scores for the OK and BAD classes, denoted as F_1-Mult. The same metric was applied to gap and source token labels. We also report F_1-scores for individual classes for completeness. We test the significance of the results using randomisation tests (Yeh, 2000) with Bonferroni correction (Abdi, 2007).

Results The results for Task 2 are summarised in Tables 9, 10, 11 and 12, ordered by the F_1-mult metric.

The number of submissions per language pair was different, which limits any conclusions that can be made with respect to general rankings of systems. The English-German and German-English tasks – Tables 9, 10 – had the most systems participating. As in previous years, results in Task1 and Task2 are correlated. In this case the same system, QEBrain, wins both tasks for these language pairs. Since some of the other systems for Task 1 where specific for sentence-level prediction, the next system in the ranking is SHEF-PT, which lags behind by a margin slightly smaller than in Task 1. Another interesting result for this year is the differences between SMT and NMT datasets. For English-German, there is a clear drop in performance from SMT to NMT. This can be due to changes in the type of errors, or size of training sets, as we discuss in Section 9.

Regarding the novel task variants of detection of gaps and source words that lead to errors, only a few teams submitted systems. The performance for these tasks is lower, but correlated with the performance of the main word-level task – prediction of target word errors. It is worth noting that the QEBrain system obtains notable performance for gap error detection, almost doubling the performance of other (few) participating systems for SMT data.

The English-Latvian and English-Czech tasks had a lower number of participants, potentially due to the lower number of resources to pre-process data and pre-train models. It is interesting to note

[12] https://github.com/Unbabel/word-level-qe-corpus-builder

Model	**Pearson** r	MAE	RMSE	Spearman ρ
SMT DATASET				
• QEBrain DoubleBi w/ BPE+word-tok (ensemble)	0.74	0.09	0.14	0.75
QEBrain DoubleBi w/ BPE-tok	0.73	0.10	0.14	0.75
UNQE	0.70	0.10	0.14	0.72
TSKQE2	0.49	0.13	0.17	0.00
SHEF-PT	0.49	0.13	0.17	0.51
TSKQE1	0.48	0.13	0.17	0.00
UTartu/QuEst+Attention	0.43	0.14	0.17	0.42
UTartu/QuEst+Att+CrEmb3	0.42	0.14	0.17	0.42
sMQE	0.40	0.19	0.22	0.40
RTM_MIX7	0.39	0.14	0.18	0.40
RTM_MIX6	0.39	0.14	0.18	0.40
SHEF-bRNN	0.37	0.14	0.18	0.38
BASELINE	0.37	0.14	0.18	0.38
uMQE	–	–	–	0.38
UAlacant**	0.39	0.18	0.23	0.39
NMT DATASET				
• UNQE	0.51	0.11	0.17	0.61
• QEBrain DoubleBi w/ BPE+word-tok (ensemble)	0.50	0.11	0.17	0.60
• QEBrain DoubleBi w/ word-tok	0.50	0.11	0.17	0.60
TSKQE1	0.42	0.14	0.18	0.00
TSKQE2	0.41	0.14	0.18	0.00
SHEF-bRNN	0.38	0.13	0.18	0.48
SHEF-PT	0.38	0.13	0.18	0.47
UTartu/QuEst+Attention	0.37	0.13	0.18	0.44
sMQE	0.37	0.21	0.24	0.44
UTartu/QuEst+Att+CrEmb3	0.37	0.13	0.18	0.44
BASELINE	0.29	0.13	0.19	0.42
uMQE	–	–	–	0.40
UAlacant**	0.23	0.21	0.26	0.24
RTM_MIX5**	0.47	0.12	0.17	0.55

Table 5: Official results of the WMT18 Quality Estimation Task 1 for the **English–German** dataset. The winning submission is indicated by a •. Baseline systems are highlighted in grey, and ** indicates late submissions that were not considered for the official ranking of participating systems.

Model	**Pearson** r	MAE	RMSE	Spearman ρ
• UNQE	0.77	0.09	0.13	0.73
• QEBrain DoubleBi w/ BPE+word-tok (ensemble)	0.76	0.10	0.13	0.73
• QEBrain DoubleBi w/ word-tok	0.75	0.10	0.14	0.72
sMQE	0.65	0.12	0.15	0.60
UTartu/QuEst+Att+CrEmb3	0.57	0.14	0.18	0.47
SHEF-PT	0.55	0.13	0.17	0.50
UTartu/QuEst+Attention	0.55	0.14	0.17	0.47
SHEF-bRNN	0.48	0.14	0.19	0.44
BASELINE	0.33	0.15	0.19	0.32
UAlacant**	0.63	0.12	0.17	0.60
RTM_MIX5**	0.54	0.13	0.17	0.49

Table 6: Official results of the WMT18 Quality Estimation Task 1 for the **German–English** dataset. The winning submission is indicated by a •. Baseline systems are highlighted in grey, and ** indicates late submissions that were not considered for the official ranking of participating systems.

Model	**Pearson** r	MAE	RMSE	Spearman ρ
SMT DATASET				
• UNQE	0.62	0.12	0.16	0.58
sMQE	0.46	0.13	0.18	0.41
UTartu/QuEst+Att+CrEmb3	0.40	0.16	0.20	0.32
UTartu/QuEst+Attention	0.40	0.15	0.19	0.32
SHEF-bRNN	0.40	0.14	0.19	0.33
SHEF-PT	0.38	0.14	0.19	0.33
BASELINE	0.35	0.16	0.19	0.35
uMQE	–	–	–	0.40
UAlacant**	0.36	0.20	0.26	0.34
RTM_MIX**	0.35	0.14	0.19	0.28
NMT DATASET				
• UNQE	0.68	0.13	0.17	0.67
sMQE	0.58	0.15	0.19	0.57
UTartu/QuEst+Att+CrEmb3	0.54	0.16	0.20	0.50
UTartu/QuEst+Attention	0.53	0.16	0.20	0.49
SHEF-PT	0.46	0.17	0.22	0.45
BASELINE	0.44	0.16	0.22	0.46
SHEF-bRNN	0.42	0.17	0.22	0.41
uMQE	–	–	–	0.54
UAlacant**	0.56	0.17	0.22	0.55
RTM_MIX**	0.54	0.16	0.20	0.50

Table 7: Official results of the WMT18 Quality Estimation Task 1 for the **English–Latvian** dataset. The winning submission is indicated by a •. Baseline systems are highlighted in grey, and ** indicates late submissions that were not considered for the official ranking of participating systems.

Model	**Pearson** r	MAE	RMSE	Spearman ρ
• UNQE	0.69	0.12	0.17	0.71
SHEF-PT	0.53	0.15	0.19	0.54
SHEF-bRNN	0.50	0.16	0.20	0.51
UTartu/QuEst+Attention	0.45	0.16	0.20	0.46
UTartu/QuEst+Att+CrEmb3	0.41	0.17	0.21	0.40
BASELINE	0.39	0.17	0.21	0.41
sMQE	0.39	0.16	0.21	0.42
uMQE	–	–	–	0.42
UAlacant**	0.44	0.18	0.23	0.46
RTM_MIX**	0.52	0.15	0.20	0.53

Table 8: Official results of the WMT18 Quality Estimation Task 1 for the **English–Czech** dataset. The winning submission is indicated by a •. Baseline systems are highlighted in grey, and ** indicates late submissions that were not considered for the official ranking of participating systems.

the general low performance of systems on the English-Latvian NMT data: all systems are tied with the baseline in terms of F_1-mult. The reimplementation of the POSTECH system shows poor results on the NMT dataset, in this case it is unable to outperform the baseline. Results for English-Czech are very similar across systems.

7 Task 3: Predicting phrase-level quality

This level of granularity was first introduced in the shared task at WMT16. The goal is to predict MT quality at the level of phrases. In the 2016 edition, the data annotation was done automatically based on post-edits, as in Task 2, but this year humans directly labelled each phrase in context.

SMT DATASET	Words in MT			GAPs in MT			Words in SRC		
Model	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**
• QEBrain DoubleBi w/ BPE+word-tok (ensemble)	0.68	0.92	0.62	–	–	–	–	–	–
QEBrain DoubleBi w/ word-tok	0.66	0.92	0.61	0.51	0.98	0.50	–	–	–
SHEF-PT	0.51	0.85	0.43	0.29	0.96	0.28	0.42	0.80	0.34
CMU-LTI	0.48	0.82	0.39	–	–	–	–	–	–
SHEF-bRNN	0.45	0.81	0.37	0.27	0.96	0.26	0.41	0.82	0.34
BASELINE	0.41	0.88	0.36	–	–	–	–	–	–
Doc2Vec	0.29	0.75	0.22	–	–	–	–	–	–
BagOfWords	0.28	0.73	0.20	–	–	–	–	–	–
UAlacant**	0.35	0.81	0.29	0.33	0.96	0.32	–	–	–
RTM**	0.33	0.88	0.29	0.26	0.98	0.25	0.17	0.86	0.14

NMT DATASET	Words in MT			GAPs in MT			Words in SRC		
Model	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**
• QEBrain DoubleBi w/ word-tok (using voting)	0.48	0.91	0.44	–	–	–	–	–	–
• QEBrain DoubleBi w/ word-tok	0.48	0.92	0.43	–	–	–	–	–	–
CMU-LTI	0.36	0.85	0.30	–	–	–	–	–	–
SHEF-bRNN	0.35	0.86	0.30	0.12	0.98	0.12	0.33	0.87	0.29
SHEF-PT	0.34	0.87	0.29	0.11	0.98	0.11	0.31	0.84	0.26
BASELINE	0.20	0.92	0.18	–	–	–	–	–	–
UAlacant**	0.23	0.86	0.19	0.12	0.98	0.12	–	–	–
RTM**	0.14	0.99	0.13	0.14	0.99	0.13	0.03	0.92	0.03

Table 9: Official results of the WMT18 Quality Estimation Task 2 for the **English–German** dataset. The winning submission is indicated by a •. Baseline systems are in grey, and ** indicates late submissions that were not considered for the official ranking of participating systems.

Model	Words in MT			GAPs in MT			Words in SRC		
	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**
• QEBrain DoubleBi w/ BPE+word-tok (ensemble)	0.65	0.92	0.60	–	–	–	–	–	–
• QEBrain DoubleBi w/ word-tok	0.65	0.92	0.59	–	–	–	–	–	–
BASELINE	0.49	0.90	0.44	–	–	–	–	–	–
SHEF-PT	0.49	0.87	0.42	0.21	0.97	0.20	0.39	0.89	0.35
CMU-LTI	0.49	0.85	0.42	–	–	–	–	–	–
SHEF-brNN	0.45	0.87	0.39	0.20	0.97	0.19	0.37	0.87	0.32
UAlacant**	0.43	0.87	0.37	0.33	0.97	0.32	–	–	–
RTM**	0.38	0.90	0.34	0.15	0.98	0.14	0.12	0.90	0.11

Table 10: Official results of the WMT18 Quality Estimation Task 2 for the **German–English** dataset. The winning submission is indicated by a •. Baseline systems are in grey, and ** indicates late submissions that were not considered for the official ranking of participating systems.

Model	Words in MT			GAPs in MT			Words in SRC		
	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**
• CMU-LTI	0.56	0.80	0.45	–	–	–	–	–	–
BASELINE	0.53	0.83	0.44	–	–	–	–	–	–
• SHEF-PT	0.56	0.80	0.44	0.17	0.98	0.17	0.49	0.80	0.39
SHEF-bRNN	0.55	0.79	0.44	0.18	0.97	0.17	0.49	0.81	0.40
UAlacant**	0.42	0.75	0.32	0.15	0.95	0.15	–	–	–
RTM**	0.53	0.83	0.44	0.11	0.98	0.10	0.32	0.80	0.26

Table 11: Official results of the WMT18 Quality Estimation Task 2 for the **English–Czech** dataset. The winning submission is indicated by a •. Baseline systems are in grey, and ** indicates late submissions that were not considered for the official ranking of participating systems.

SMT DATASET	Words in MT			GAPs in MT			Words in SRC		
Model	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**
• SHEF-PT	0.42	0.87	0.36	0.14	0.97	0.14	0.35	0.86	0.30
• SHEF-bRNN	0.41	0.86	0.35	0.12	0.98	0.11	0.36	0.86	0.31
BASELINE	0.38	0.91	0.34	–	–	–	–	–	–
CMU-LTI	0.22	0.85	0.19	–	–	–	–	–	–
UAlacant**	0.27	0.82	0.22	0.11	0.96	0.11	–	–	–
RTM**	0.37	0.90	0.33	0.13	0.99	0.13	0.12	0.89	0.11

NMT DATASET	Words in MT			GAPs in MT			Words in SRC		
Model	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**
• CMU-LTI	0.52	0.83	0.43	–	–	–	–	–	–
BASELINE	0.49	0.86	0.42	–	–	–	–	–	–
• SHEF-PT	0.52	0.81	0.42	0.13	0.97	0.13	0.44	0.81	0.36
• SHEF-bRNN	0.50	0.83	0.42	0.12	0.94	0.11	0.44	0.80	0.36
UAlacant**	0.45	0.80	0.36	0.17	0.95	0.16	–	–	–
RTM**	0.43	0.85	0.37	0.08	0.98	0.08	0.20	0.84	0.17

Table 12: Official results of the WMT18 Quality Estimation Task 2 for the **English–Latvian** dataset. The winning submission is indicated by a •. Baseline systems are in grey, and ** late submissions that were not considered for the official ranking of participating systems.

Labels We used the phrase segmentation produced by the SMT decoder which generated the translations for the dataset. The phrases were annotated for errors using four classes: 'OK', 'BAD' – the phrase contain one or more errors, 'BAD_word_order' – the phrase is in an incorrect position in the sentence, and 'BAD_omission' – a word is missing before/after a phrase. This task in further subdivided in two subtasks: word-level prediction (Task3a), and phrase-level prediction (Task3b).

The data for Task3a propagates the annotation of each phrase to its words, and thus uses word-level segmentation for both source and machine-translated sentences, such that the task can be addressed as a word-level prediction task. In other words, all tokens in the target sentence are labelled according to the label of the phrase they belong to. Therefore, if the phrase is annotated as either 'OK', 'BAD' or 'BAD_word_order', all tokens (and gap tokens) within that phrase are labelled as either 'OK', 'BAD' or 'BAD_word_order'. To annotate omission errors, a gap token is inserted after each token and at the start of the sentence.

The data for Task3b has phrase-level segmentation with the labels assigned by the human annotator to each phrase. A gap token is inserted after each phrase and at the start of the sentence. The gap is labelled as follows: 'OK' or 'BAD_omission', where the latter indicates that one or more words are missing.

Evaluation Similarly to Task 2, our primary metric for predictions at word-level (Task3a) is the multiplication of the F_1 scores of the OK and BAD classes, F_1-Mult, while for predictions at phrase-level (Task3b), our primary metric is the phrase-level version of F_1-Mult. The same metrics were applied to gap and source token labels for both sub-tasks, along with F_1 scores for individual classes for completeness. We also report F_1 score for BAD_word_order labels on the target tokens for Task3b. We computed statistical significance of the results using randomised test with Bonferroni correction, as in Task 2.

Results The results of the phrase-level task are given in Tables 13 (Task3a) and 14 (Task3b), ordered by the F_1-Mult metric.

Comparing the results for Task3a with the results on German-English for Task 2 (Table 10), it can be observed a general degradation of the F_1 score on the BAD class, including for the baseline system. We attribute this phenomenon to the way the data for this task was created: for Task 2, the token labels were produced from post-editing, where each word was labelled independently from each others; while for this task, the token labels are deduced from a labelling at more coarse level (phrase), i.e. where words were not considered as individual tokens. Consequently, words that would be considered as correct during post-editing are here labelled as BAD, like to BAD phrase they belong to. The only two official submissions to this subtask (SHEF-PT and SHEF-bRNN) slightly outperform the baseline system, nevertheless without a statistically significant difference.

For the phrase-level predictions, the baseline system remains ahead by a significant margin of the only two official submissions, both from the University of Sheffield (SHEF-ATT-SUM and SHEF-PT). The overall performance in predicting phrases that are in incorrect position in a sentence (i.e. BAD_word_order) shows that this problem remains a very challenging task, as none of the submissions were able to obtain competitive F_1 score.

8 Task 4: Predicting document-level QE

This task consists in estimating a document-level quality score according to the amount of minor, major, and critical errors present in the translation. The predictions are compared to a ground-truth obtained from annotations produced by crowd-sourced human translators from Unbabel.[13]

Labels The data was annotated for errors at the word level using a fine-grained error taxonomy – Multidimensional Quality Metrics (MQM) (Lommel et al., 2014) – similar to the one used in (Sanchez-Torron and Koehn, 2016). MQM is composed of three major branches: accuracy (the translation does not accurately reflect the source text), fluency (the translation affects the reading of the text) and style (the translation has stylistic problems, like the use of a wrong register). These branches include more specific issues lower in the hierarchy. Besides the identification of an error and its classification according to this typology (by applying a specific tag), the errors receive a severity scale that reflects the impact of each error on the overall meaning, style, and fluency of the translation. An error can be *minor* (if it does

[13] http://www.unbabel.com.

Model	Words in MT			GAPs in MT			Words in SRC		
	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**
• SHEF-PT	0.33	0.83	0.28	0.27	0.88	0.24	0.50	0.81	0.41
• SHEF-bRNN	0.33	0.82	0.27	0.26	0.88	0.23	0.49	0.79	0.39
BASELINE	0.27	0.91	0.25	–	–	–	–	–	–
RTM**	0.16	0.90	0.15	0.10	0.94	0.10	0.10	0.84	0.08

Table 13: Official results of the WMT18 Quality Estimation Task 3a (word-level) for the **German–English** dataset. The winning submission is indicated by a •. Baseline system is in grey, and ** indicates late submissions that were not considered for the official ranking of participating systems.

Model	Words in MT				GAPs in MT			Words in SRC		
	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD_w_order	F_1-BAD	F_1-OK	**F_1-mult**	F_1-BAD	F_1-OK	**F_1-mult**
BASELINE	0.39	0.92	0.36	0.02	–	–	–	–	–	–
• SHEF-ATT-SUM	0.29	0.76	0.22	0.11	0.10	0.94	0.10	–	–	–
SHEF-PT	0.23	0.81	0.18	0.08	0.11	0.93	0.10	–	–	–
RTM**	0.27	0.92	0.24	0.04	0.05	0.98	0.05	0.10	0.90	0.09

Table 14: Official results of the WMT18 Quality Estimation Task 3b (phrase-level) for the **German–English** dataset. The winning submission is indicated by a •. Baseline system is in grey, and ** indicates late submissions that were not considered for the official ranking of participating systems.

not lead to a loss of meaning and it doesn't confuse or mislead the user), *major* (if it changes the meaning) or *critical* (if it changes the meaning and carry any type of implication, or could be seen as offensive).

Document-level scores were then generated from the word-level errors and their severity using the method described in Sanchez-Torron and Koehn (2016, footnote 6). Namely, denoting by n the number of words in the document, and by n_{min}, n_{maj}, and n_{cri} the number of annotated minor, major, and critical errors, the final quality scores were computed as:

$$\text{MQM Score} = 1 - \frac{n_{\text{min}} + 5n_{\text{maj}} + 10n_{\text{cri}}}{n} \quad (1)$$

Note that MQM values can be negative if the total severity exceeds the number of words.

Evaluation Submissions are evaluated as in Task 1 (see Section 5), in terms of Pearson's correlation r between the true and predicted document-level scores.

Results The results of the document-level task are shown in Table 15. Due to the different numeric range, only the Pearson correlation scores are comparable to those of Task1. Comparing with the results for Task 1, it can be observed that the baseline system already obtains very high correlation. The neural model SHEF-PT-indomain outperforms the baseline by a modest margin, compared to the results obtained in Task 1.

Model	**Pearson** r	MAE
• SHEF-PT-indomain	0.53	0.56
BASELINE	0.51	0.56
SHEF-mtl-bRNN	0.47	0.56
SHEF-mtl-PT-indomain**	0.52	0.57
RTM_MIX1**	0.11	0.58

Table 15: Official results of the WMT18 Quality Estimation Task 4 for the **English–French** dataset. The winning submission is indicated by a •. Baseline system is in grey, and ** indicates late submissions that were not considered for the official ranking of participating systems.

9 Discussion

In what follows, we discuss the main findings of this year's shared task based on the goals we had previously identified for it.

Performance of QE approaches on the output of neural MT systems. As previously mentioned, some of the data used for Tasks 1 and 2 is translated by both an SMT and an NMT system: the English-German and English-Latvian data. In Task 1, for English-German, the numbers of translations in the QE training data from the two systems are very different ($26,273$ for SMT and $13,442$ for NMT) and thus no direct comparison can be made. This shows that the NMT system was of much higher quality than the SMT one, producing many more sentences that led to HTER=0. Of the sentences that remained, the average NMT quality in the training data is still higher: HTER=0.154 versus 0.253 for SMT. From the results, the top systems do considerably better on the SMT data (r=0.74 for SMT vs r=0.51 for NMT translations). This difference is also noticeable for the baseline system (r=0.37 for SMT vs r=0.29 for NMT translations). This could however be because of the difference in number of samples and/or significant differences in distributions of HTER scores in the two datasets. It is worth pointing out that the winning submissions are the same for both SMT and NMT translation: QEBrain and UNQE. In fact, QE system are ranked very similarly for the two types of translation.

For English-Latvian, the number of NMT and SMT QE training sentences is similar ($12,936$ for NMT and $11,251$ for SMT). Their average HTER scores is also more comparable: 0.278 for NMT and 0.215 for SMT. The difference in QE system performance for this language pair is not as marked, but the trend is inverted when compared to English-German: QE systems do better on the NMT data (the top systems, UNQE, achieves r=0.62 for SMT vs r=0.68 for NMT translations, while the baseline achieves r=0.44 for SMT vs r=0.35 for NMT translations), This could be because of the lower differences in the distribution of HTER scores in both sets. The ranking of QE systems is exactly the same for both SMT and NMT translations.

In both cases, it is important to note that even though the initial datasets contained exactly the same source sentences for SMT and NMT, the sentences in the two final versions of the datasets for each language are not all the same, i.e. some NMT sentences may have gotten filtered for having HTER=0 while their SMT counterparts did

not, and vice-versa. The main finding is that QE models seem to be robust to different types of translation, since their rankings are the same across datasets.

For Task 2, the trend is similar: QE systems for English-German also perform better on SMT translations than on NMT translations (F_1-Mult=0.62 for SMT vs F_1-Mult=0.44 for NMT), and the inverse is observed for English-Latvian (F_1-Mult=0.36 for SMT vs F_1-Mult=0.43 for NMT). The ranking of QE systems for the two types of translations differs more than for Task 1, especially for English-Latvian.

Task 4 uses NMT output only and it is hard to make any conclusions about whether the performance of the systems is good enough because this is the first time this task is organised. Generally speaking, this task proved hard, with the baseline system performing as well or better than the other submissions.

Predictability of missing words in the MT output. Only a subset of the systems that participated in Task 2 submitted results for missing word detection. From the results obtained it seems clear that while this task is more difficult than target word error detection, high scores could be attained for the SMT data. Due to the small number of submitted systems, it is unclear whether or not gap detection is more difficult for NMT data.

Predictability of source words that lead to errors in the MT output. Only a small set of teams submitted predictions for source words. From the submitted results, it can be observed that prediction of source words related to errors is a harder problem than detecting errors in the target language. This may be due to the fact that there may be more ambiguity with regards to which words should be related to errors in the target. In other words, in some cases a source word in a given context leads to incorrect translations, while in other cases the same source word in the same context will not lead to errors.

Effectiveness of manually assigned labels for phrases. With only one official (and one late) submission to the phrase-level QE task this year, it is hard to conclude whether having manual labels makes the task harder (although the baseline system performs as well as in the last edition), or whether the reason lies in the design of the neural models, which may not be suitable for this task.

Quality prediction for documents from errors annotated at word-level with added severity judgements. Since this is a new task and not many systems were submitted. Results show however that it is possible to attain Pearson correlation scores that are comparable with those of sentence-level post-editing effort prediction. The performance gap between the neural model and the SVM baseline is smaller than in Task 1, which may be an indication for further potential gains using new deep learning architectures tailored for document-level.

Utility of additional evidence To investigate the utility of detailed information logged during post-editing, we offered to participants other sources of information: post-editing time, keystrokes, and actual edits. Surprisingly, no participating system requested these additional labels, and therefore this remains an open question.

10 Conclusions

This year's edition of the QE shared task was the largest ever organised in many respects: number of tasks, number of languages, variety of tasks (three granularity levels), types of annotation (derived from post-editing or manual, source or target), and number of samples annotated.

Over the years, we have attempted to find a balance between keeping the shared task as close as possible to previous editions – so as to make some form of comparison across years possible – and proposing new tasks and new interesting challenges – so as to keep up to date with new developments in the field, such as neural machine translations. We believe the current set of tasks covers a broad enough range of challenges that are far from solved, such as improving performance given smaller sets of instances, predicting source words that lead to errors, predicting gaps, use of additional evidence from post-editing, etc.

In order to allow for future benchmarking on a 'blind' basis without access to the gold standard labels, we have set up CodaLab competitions that will remain open after this shared task. Any team can register and submit any number of systems (limited to five submissions per day per task and language pair) and get immediate feedback through the official evaluation metrics, as well as comparison to top submissions from other teams (on the leaderboard). Each team's best submission per task and language pair will feature on the

leaderboard. The submission pages for each task are as follows, where languages and task variants are frames as 'phases':

- Sentence level: https://competitions.codalab.org/competitions/19316
- Word level: https://competitions.codalab.org/competitions/19306
- Phrase level: https://competitions.codalab.org/competitions/19308
- Document level: https://competitions.codalab.org/competitions/19309

Acknowledgments

The data and annotations collected for Tasks 1, 2 and 3 was supported by the EC H2020 QT21 project (grant agreement no. 645452). The shared task organisation was also supported by the QT21 project, national funds through Fundação para a Ciência e Tecnologia (FCT), with references UID/CEC/50021/2013 and UID/EEA/50008/2013, and by the European Research Council (ERC StG DeepSPIN 758969). We would also like to thank Julie Belião and the Unbabel Quality Team for coordinating the annotation of the dataset used in Task 4.

References

Hervé Abdi. 2007. The bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3:103–107.

Wilker Aziz, Sheila Castilho Monteiro de Sousa, and Lucia Specia. 2012. Pet: a tool for post-editing and assessing machine translation. In *Eighth International Conference on Language Resources and Evaluation*, LREC, pages 3982–3987, Istanbul, Turkey.

Prasenjit Basu, Santanu Pal, and Sudip Kumar Naskar. 2018. Keep it or not: Word level quality estimation for post-editing. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Brussels, Belgium. Association for Computational Linguistics.

Ergun Biçici. 2017. Predicting translation performance with referential translation machines. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 540–544, Copenhagen, Denmark. Association for Computational Linguistics.

Ergun Biçici. 2018. Rtm results for predicting translation performance. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Brussels, Belgium. Association for Computational Linguistics.

Ergun Bicici. 2013. Referential translation machines for quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 343–351, Sofia, Bulgaria. Association for Computational Linguistics.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Eighth Workshop on Statistical Machine Translation*, WMT, pages 1–44, Sofia, Bulgaria.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Ninth Workshop on Statistical Machine Translation*, WMT, pages 12–58, Baltimore, Maryland.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012.

Findings of the 2012 workshop on statistical machine translation. In *Seventh Workshop on Statistical Machine Translation*, WMT, pages 10–51, Montréal, Canada.

Melania Duma and Wolfgang Menzel. 2018. The benefit of pseudo-reference translations in quality estimation of mt output. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Brussels, Belgium. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Thierry Etchegoyhen, Eva Martínez Garcia, and Andoni Azpeitia. 2018. Supervised and unsupervised minimalist quality estimators: Vicomtech's participation in the wmt 2018 quality estimation task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Brussels, Belgium. Association for Computational Linguistics.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517. International World Wide Web Conferences Steering Committee.

Junjie Hu, Wei-Cheng Chang, Yuexin Wu, and Graham Neubig. 2018. Contextual encoding for translation quality estimation. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Brussels, Belgium. Association for Computational Linguistics.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018a. DeepQuest: a framework for neural-based quality estimation. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics: Technical Papers*, Santa Fe, New Mexico.

Julia Ive, Carolina Scarton, Frédéric Blain, and Lucia Specia. 2018b. Sheffield submissions for the wmt18 quality estimation shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Brussels, Belgium. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Julia Kreutzer, Shigehiko Schamoni, and Stefan Riezler. 2015. QUality Estimation from ScraTCH (QUETCH): Deep Learning for Word-level Translation Quality Estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 297–303, Lisboa, Portugal. Association for Computational Linguistics.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196.

Maoxi Li, Qingyu Xiang, Zhiming Chen, and Mingweng Wang. 2018. A unified neural network for quality estimation of machine translation. *IEICE Trans. Information and Systems*, E101-D(9).

Varvara Logacheva, Chris Hokamp, and Lucia Specia. 2016. Marmot: A toolkit for translation quality estimation at the word level. In *Tenth International Conference on Language Resources and Evaluation*, LREC, pages 3671–3674, Portoroz, Slovenia.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0(12):455–463.

Ngoc Quang Luong, Laurent Besacier, and Benjamin Lecouteux. 2014. Word confidence estimation for smt n-best list re-ranking. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 1–9, Gothenburg, Sweden. Association for Computational Linguistics.

André F. T. Martins, Ramón Astudillo, Chris Hokamp, and Fabio Kepler. 2016. Unbabel's participation in the wmt16 word-level translation quality estimation shared task. In *Proceedings of the First Conference on Machine Translation*, pages 806–811, Berlin, Germany. Association for Computational Linguistics.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM.

Sylvain Raybaud, David Langlois, and Kamel Smaili. 2011. "this sentence is wrong." detecting errors in machine-translated sentences. *Machine Translation*, 25(1):1–34.

Felipe Sánchez-Martíínez, Miquel Esplà-Gomis, and Mikel L. Forcada. 2018. Ualacant machine translation quality estimation at wmt 2018: a simple approach using phrase tables and feed-forward neural networks. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Brussels, Belgium. Association for Computational Linguistics.

Marina Sanchez-Torron and Philipp Koehn. 2016. Machine translation quality and post-editor productivity. *AMTA 2016, Vol.*, page 16.

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Viviven Macketanz, Inguna Skadina, Matteo Negri, , and Marco Turchi. 2017. Translation quality and productivity: A study on rich morphology languages. In *Machine Translation Summit XVI*, pages 55–71, Nagoya, Japan.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba submission for wmt18 quality estimation task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Brussels, Belgium. Association for Computational Linguistics.

Elizaveta Yankovskaya, Andre Tattar, and Mark Fishel. 2018. Quality estimation with force-decoded attention and cross-lingual embeddings. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Brussels, Belgium. Association for Computational Linguistics.

Alexander Yeh. 2000. More Accurate Tests for the Statistical Significance of Result Differences. In *Coling-2000: the 18th Conference on Computational Linguistics*, pages 947–953, Saarbrücken, Germany.