Sheffield Submissions for the WMT18 Quality Estimation Shared Task

Julia Ive[1], **Carolina Scarton**[2], **Frédéric Blain**[2] and **Lucia Specia**[2]
[1]IoPPN, Kings College London, UK
[2]DCS, University of Sheffield, UK
julia.ive@kcl.ac.uk
{c.scarton, f.blain, l.specia}@sheffield.ac.uk

Abstract

In this paper we present the University of Sheffield submissions for the WMT18 Quality Estimation shared task. We discuss our submissions to all four sub-tasks, where ours is the only team to participate in all language pairs and variations (37 combinations). Our systems show competitive results and outperform the baseline in nearly all cases.

1 Introduction

Quality Estimation (QE) predicts the quality of Machine Translation (MT) when automatic evaluation or human assessment is not possible (typically at system run-time). QE is mainly addressed as a supervised Machine Learning problem with QE models trained using labelled data. These labels differ for different tasks, for example, binary labels for fine-grained predictions (e.g. OK/BAD for words or phrases) and continuous measurements of quality for coarse-grained levels (e.g. HTER (Snover et al., 2006) for sentences).

For this year's shared task, post-edited (PE) and manually annotated data were provided. They cover four levels of predictions: sentence-level (task 1), word-level (task 2), phrase-level (task 3) and document-level (task 4), over five language pairs: English into German, Latvian, Czech and French, as well as German-English. For the first time, these data contain translations produced by neural MT (NMT) systems. Such translations are known to be more fluent but less adequate (Toral and Sánchez-Cartagena, 2017).

For tasks 2 and 3, this year's edition introduces a new task variant of predicting missing words in the translations. Thus two additional prediction types are required: (i) binary labels for gaps in the translation to indicate whether one or more tokens are missing from a certain position, and (ii) binary labels for words in source sentences to indicate which of these words lead to incorrect words in the translations.

We participated with two different systems, both available in the DeepQuest[1] toolkit (Ive et al., 2018):

- **SHEF-PT:** an in-house re-implementation of the POSTECH system (Kim et al., 2017b), and
- **SHEF-bRNN:** a bidirectional recurrent neural network (bRNN) system.

We participated in all sub-tasks and submitted a total of 74 predictions (37 per system).

2 Systems Description

Our light-weight neural QE approach is based on simple encoders and requires no pre-training (bRNN). We compare its performance to the performance of our re-implementation of the state-of-the-art neural QE approach of Kim et al. (2017a,b) (POSTECH), which uses a complex architecture and requires resource-intensive pre-training.

2.1 Architecture

Following current best practices in neural sequence-to-sequence modelling (Sutskever et al., 2014; Bahdanau et al., 2015), our bRNN approach employs encoders using recurrent neural networks (RNNs). Encoders encode input into an internal representation used to make classification decisions. bRNN representations at a given level rely on representations from more fine-grained levels (i.e. sentences for document, and words for phrase and sentence).

bRNN uses two bi-directional RNNs to learn the representation of the <source, MT> sentence pair. Source and MT RNNs are trained independently.

[1]https://sheffieldnlp.github.io/deepQuest

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 807–813

https://doi.org/10.18653/v1/W18-64091

The two representations are then combined via concatenation. For word-level QE, those representations (sequences of hidden states h_j associated with words) can be used directly to make classification decisions. A sentence vector is a weighted sum of word vectors as generated by an attention mechanism. Another output layer takes this sentence vector as input and produces real-value sentence-level quality scores.

For phrase-level QE, we have modified the architecture described above. It takes a three-dimensional MT input (batch length $\times$ sentence length in phrases $\times$ phrase length in words).[2] Concatenation of source and MT sentence representations, as performed in our word- and sentence-level architecture, will require source inputs to be three-dimensional as well. However, as the phrase alignments are not provided with the task, three-dimensional source inputs can not be formed without an additional approximation.[3] Instead, we follow best practices of NMT (Bahdanau et al., 2015) and implement its standard encoder-decoder architecture. The encoder creates source representations using a bidirectional RNN, at each timestep the decoder produces a word representation taking into account not only the previously produced representations, but also the sum of source word representations weighted by an attention mechanism.[4] This process can be interpreted as defining word alignments: the resulting decoder representations contain information on both MT words and respective parts of the source attended at each timestep. Each phrase representation can be computed out of word vectors: average, maximum, sum, etc. The resulting representations are provided to the output layer, as illustrated in Figure 1.

Our document-level framework is a wrapper over sentence QE approaches. It uses a bidirectional RNN to summarize sentence-level representations as document-level representations used for regression.

More details on the architecture and implementation of our sentence and document-level models can be found in Ive et al. (2018).

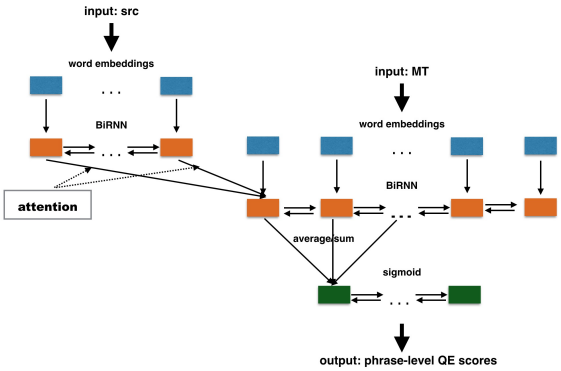

Figure 1: bRNN phrase-level QE architecture.

2.2 Implementation Details

To train POSTECH's predictor, we used the corresponding parts of the in-domain corpora provided by the organisers for the corresponding languages ($\approx$ 2M sentences were selected randomly per language pair). The only exception was EN-LV for which we had less than 2M sentences in the corpus. Therefore, we combined the in-domain corpus with the Europarl (version 8)[5] and EMEA corpus.[6] This totaled in 1,241,615 EN-LV sentences.

For the word and phrase-level tasks, we tackled prediction of MT error tags, source tags and MT gaps separately. For predicting source tags, we built models by swapping source and MT inputs. POSTECH's predictors were then trained with swapped source and target inputs. For predicting gaps, we added a dummy word at the beginning of each MT sentence to match the count of gap tags per line.

We experimented with phrase-level representations and created them by computing the sum or the average of composing word vectors. To optimise the usage of computational resources, in each experiment we fixed the size of a phrase in words to the upper quartile of the respective distribution in the training data.

For the document-level QE, we experimented with sentence-level representations coming from both bRNN and POSTECH architectures.

For our POSTECH-based document-level models, we experimented with predictors trained on a

[2]Note that other architectural choices may lead to, for instance, two-dimensional inputs (batch length $\times$ phrase length in words). A representation of each MT phrase may be created without taking the rest of the translated sentence into account.

[3]For instance, we may assume that translation of a phrase relies on the whole source sentence. Thus, a three-dimensional input can be formed by simply repeating each source sentence along the second axis to match respective counts of phrases in each MT sentence.

[4]Note that Jhaveri et al. (2018) also use this architecture for sentence-level QE.

[5]http://www.statmt.org/wmt17/translation-task.html

[6]http://opus.nlpl.eu/EMEA.php

	SHEF-PT			SHEF-bRNN			Baseline		
	r	MAE	ρ	r	MAE	ρ	r	MAE	ρ
EN-DE – SMT	**0.487**	**0.132**	**0.510**	0.366	0.139	0.378	0.365	0.140	0.381
EN-DE – NMT	0.377	0.131	0.468	**0.381**	0.130	**0.480**	0.287	**0.129**	0.420
EN-LV – SMT	0.375	0.141	0.329	**0.396**	**0.138**	0.332	0.353	0.155	**0.348**
EN-LV – NMT	**0.463**	0.166	0.446	0.421	0.172	0.409	0.444	**0.163**	**0.458**
EN-CS	**0.533**	**0.150**	**0.537**	0.501	0.157	0.506	0.394	0.165	0.414
DE-EN	**0.554**	**0.130**	**0.501**	0.482	0.143	0.443	0.332	0.151	0.325

Table 1: Evaluation of our systems for task 1 on the test set. We show scores of Pearson's r correlation, MAE and Spearman's ρ correlation.

part of the English–French Europarl (version 7),[7] as well as on an in-domain corpus (described in Section 3.4). As mentioned before, our document-level QE system is a modular architecture wrapping over any sentence-level QE model. We took advantage of this modularity and also attempted multi-task learning (MTL). We pre-trained the weights of sentence-level modules (both `bRNN` and `POSTECH`) to predict Multidimensional Quality Metrics (MQM)[8] scores for sentences (more details in Section 3.4).

3 Tasks Participation

The four QE tasks correspond to different levels of quality prediction: sentence-level (task 1), word-level (task 2 and 3a), phrase-level (task 3b) and document-level (task 4). For each prediction level, different language pairs and system outputs are provided. Below we provide a detailed description of the datasets together with the results for our submitted systems for each of these tasks.

3.1 Task 1: Sentence-level QE

Four language pairs are available for sentence-level scoring and ranking:

- EN-DE: sentences on the IT domain, with MT from either an SMT (26,273 training / 1,000 development / 1,000 test) or an NMT (13,442 training / 1,000 development / 1,000 test) system,
- EN-LV: sentences on the life sciences domain, with MT from either an SMT (11,251 training / 1,000 development / 1,000 test) or an NMT (12,936 training / 1,000 development / 1,000 test) system,
- EN-CS: sentences on the IT domain, with MT from an SMT system (40,254 training / 1,000 development / 1,000 test), and
- DE-EN: sentences on the life sciences domain, with MT from an SMT system (25,963 training / 1,000 development / 1,000 test).

In summary, there are six data setting variants and the quality score for prediction is HTER in all of them. For each variant in this task we submitted two systems: SHEF-PT and SHEF-bRNN. For the ranking evaluation, we rank sentences using the predicted HTER outputted by our systems.

Following the shared task setup, Pearson's r correlation coefficient is used as the primary evaluation metric for the scoring task (with Mean Absolute Error – MAE – as the secondary metric), whilst Spearman's ρ rank correlation coefficient is used as metric for the ranking task. The task baseline systems are Support Vector Machine (SVM) models trained with 17 baseline features from QuEst++ (Specia et al., 2015).

We show the official results in Table 1. Both our systems outperform the baseline for all the language pairs according to the main evaluation metric (r). SHEF-bRNN is better than SHEF-PT only for EN-DE – NMT and EN-LV – SMT. These may be cases where `bRNN` is able to better capture the fluency of high-quality MT by encoding it directly as sequences rather than assessing it word for word as `POSTECH`. On the official development set,[9] EN-DE – NMT and EN-LV – SMT translations have the best overall quality (on average HTER=0.17 versus HTER=0.28 for the rest of the systems).

3.2 Task 2: Word-level QE

Task 2 uses the same datasets as task 1. Target words are assigned a binary label (OK or BAD) based on the alignments between MT and post-edits extracted by the TER tool. In this year's edition, the organisers have also proposed the predic-

[7] http://www.statmt.org/wmt15/translation-task.html

[8] http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

[9] The organisers have not provided the gold labels for the test set.

TRG words prediction									
	SHEF-PT			SHEF-bRNN			Baseline		
	F1 BAD	F1 OK	F1-MULT	F1 BAD	F1 OK	F1-MULT	F1 BAD	F1 OK	F1-MULT
EN-DE – SMT	**0.508**	0.846	**0.430**	0.453	0.811	0.367	0.412	**0.882**	0.363
EN-DE – NMT	0.335	0.869	0.291	**0.351**	0.863	**0.303**	0.197	**0.918**	0.181
EN-LV – SMT	**0.416**	0.869	**0.361**	0.409	0.860	0.351	0.381	**0.905**	0.345
EN-LV – NMT	**0.519**	0.809	0.420	0.503	0.828	0.416	0.487	**0.864**	**0.421**
EN-CS	**0.556**	0.796	0.443	0.554	0.792	0.439	0.534	**0.834**	**0.445**
DE-EN	**0.485**	0.874	0.424	0.446	0.871	0.389	**0.485**	**0.902**	**0.437**
SRC words prediction									
	SHEF-PT			SHEF-bRNN			Baseline		
	F1 BAD	F1 OK	F1-MULT	F1 BAD	F1 OK	F1-MULT	F1 BAD	F1 OK	F1-MULT
EN-DE – SMT	**0.422**	0.799	0.337	0.414	**0.821**	**0.340**	-	-	-
EN-DE – NMT	0.314	0.841	0.264	**0.330**	**0.865**	**0.286**	-	-	-
EN-LV – SMT	0.351	**0.859**	0.302	**0.357**	0.857	**0.306**	-	-	-
EN-LV – NMT	**0.444**	**0.814**	**0.361**	**0.444**	0.800	0.355	-	-	-
EN-CS	**0.493**	0.799	0.394	0.490	**0.811**	**0.398**	-	-	-
DE-EN	**0.392**	**0.887**	**0.348**	0.366	0.875	0.320			
Gaps prediction									
	SHEF-PT			SHEF-bRNN			Baseline		
	F1 BAD	F1 OK	F1-MULT	F1 BAD	F1 OK	F1-MULT	F1 BAD	F1 OK	F1-MULT
EN-DE – SMT	**0.294**	**0.962**	**0.282**	0.271	0.955	0.259	-	-	-
EN-DE – NMT	0.110	0.984	0.108	**0.121**	**0.985**	**0.119**	-	-	-
EN-LV – SMT	**0.141**	0.968	**0.136**	0.118	**0.975**	0.115	-	-	-
EN-LV – NMT	**0.130**	**0.965**	**0.126**	0.119	0.944	0.113	-	-	-
EN-CS	0.171	**0.977**	0.167	**0.179**	0.972	**0.174**	-	-	-
DE-EN	**0.210**	**0.970**	**0.204**	0.200	0.966	0.193	-	-	-

Table 2: Evaluation of our systems for task 2 on the test set. We show scores of F1-MULT, F1 for the OK class and F1 for the BAD class.

tion of gaps and source words quality. According to the TER alignment, all source words aligned to a target word will receive the same tag as the target word. For annotating gaps, a gap tag is placed after each token and in the beginning of the sentence. A gap tag will be BAD if one or more words were expected to appear in the gap, and OK otherwise.

Task 2 has 18 variants, for each of them we again submitted two systems: SHEF-PT and SHEF-bRNN.

The primary evaluation metric of task 2 is F1-MULT: multiplication of F1-scores for the OK and BAD classes. F1-scores of OK and BAD classes are used as secondary metrics. The baseline system for the target word predictions is a Conditional Random Fields (CRF) model trained with word-level baseline features from the Marmot (Logacheva et al., 2016) toolkit. There are no baseline systems for the prediction of gaps or source word issues.

Table 2 shows the official results. For prediction of target words, SHEF-PT is the best for EN-DE – SMT, EN-LV – SMT and EN-LV – NMT. SHEF-bRNN is the best for EN-DE – NMT. This confirms our previous conclusion that `bRNN` better captures the fluency of high-quality MT (cf. Section 3.1). For source words and gaps prediction, SHEF-bRNN and SHEF-PT show similar performance across language pairs.

To get a closer insight into the performance of our models, we manually analysed results for the official EN-DE – SMT/NMT development sets. For those two systems either SHEF-PT, or SHEF-bRNN performs the best respectively. Our observations suggest that, because of pre-training, SHEF-PT better captures SMT adequacy (cf. examples in Table 3; the term "screen readers" is correctly translated by the SMT system into German as "Bildschirmlesehilfen" and correctly marked as OK by SHEF-PT, but incorrectly marked as BAD by SHEF-bRNN). SHEF-bRNN better captures NMT fluency: e.g. only the word "Transparenzeffekte" correctly marked as BAD from the first part of the NMT translation in Table 3 vs. the context of this word marked as BAD by SHEF-PT.

3.3 Task 3: Phrase-level QE

This task considers a subset of the English-German SMT data from task 1 (Section 3.1). Here, the MT output has been manually anno-

SRC	to make your content accessible to screen readers , avoid using these modes .
PE	um den Inhalt für Bildschirmlesehilfen zugänglich zu machen , vermeiden Sie diese Modi .
SMT	um den Inhalt für Bildschirmlesehilfen zugänglich machen , vermeiden Sie diese Modi .
gold	OK OK OK OK OK OK OK OK OK OK OK OK OK
PT	OK OK OK OK OK OK OK OK OK OK OK OK OK
bRNN	OK OK OK OK **BAD BAD BAD BAD BAD BAD BAD BAD** OK
SRC	besides applying transparency effects to single objects , you can apply them to groups .
PE	Sie können Transparenzeffekte nicht nur auf einzelne Objekte , sondern auch auf Gruppen anwenden .
NMT	Sie können nicht nur Transparenzeffekte auf einzelne Objekte anwenden , sondern auch auf Gruppen anwenden .
gold	OK OK OK OK BAD OK OK OK BAD OK OK OK OK OK OK OK
PT	**BAD BAD BAD BAD** BAD **BAD** OK OK **OK** OK **BAD BAD BAD BAD** OK **BAD**
bRNN	OK OK OK OK BAD OK OK OK **OK** OK OK **BAD** OK OK OK OK

Table 3: Examples of prediction errors for task 2 on the EN-DE – SMT/NMT development sets

	SHEF-PT			SHEF-bRNN			Baseline		
	F1 BAD	F1 OK	F1-MULT	F1 BAD	F1 OK	F1-MULT	F1 BAD	F1 OK	F1-MULT
TRG words	**0.3338**	0.8250	**0.2754**	0.3253	0.8235	0.2679	0.2714	**0.9099**	0.2469
Gaps	**0.2730**	0.8775	**0.2396**	0.2631	**0.8785**	0.2312	-	-	-
SRC words	**0.5048**	**0.8137**	**0.4108**	0.4920	0.7916	0.3895	-	-	-

Table 4: Evaluation of our systems for task 3a on the test set. We show scores of F1-MULT, F1 for the OK class and F1 for the BAD class.

SHEF-PT				
	F1 BAD	F1 OK	F1-MULT	F1 BAD_w_o
TRG phrases	0.2294	0.8059	0.1849	0.0794
Gaps	**0.1073**	0.9349	**0.1003**	-
SHEF-ATT-SUM				
	F1 BAD	F1 OK	F1-MULT	F1 BAD_w_o
TRG phrases	0.2881	0.7614	0.2194	**0.1146**
Gaps	0.1028	**0.9416**	0.0968	-
Baseline				
	F1 BAD	F1 OK	F1-MULT	F1 BAD_w_o
TRG phrases	**0.3919**	**0.9152**	**0.3584**	0.0194
Gaps	-	-	-	-

Table 5: Evaluation of our systems for task 3b on the test set. We show scores of F1-MULT, F1 for the OK class, F1 for the BAD class and F1 for the BAD_word_order class.

tated at the phrase level with four labels: OK, BAD, BAD_word_order and BAD_omission, with the phrase boundaries defined by the SMT decoder. The last two labels are new to this task. They indicate whether a phrase is in an incorrect position in the sentence, or one or more word(s) are missing in a certain position, respectively. The subtasks of predicting gaps and source phrases quality were proposed similarly to task 2 (cf. Section 3.2).

The subtask data are provided with word-level segmentation. Task 3 is therefore divided into two subtasks 3a and 3b, for word- and phrase-level predictions, respectively.

Task3a – word-level prediction Word-level labels have been produced as follows: each word has been labelled according to the phrase it belongs to (i.e. as either OK, BAD or BAD_word_order); gaps have been labelled as either OK or BAD_omission. The evaluation metrics for this subtask are similar to task 2.

The official results are reported in Table 4. Our two systems outperform the baseline for the target words prediction, while there are no other results for gaps and source words predictions.

Task3b – phrase-level prediction In addition to the usual binary labels (OK and BAD), this subtask considers the BAD_word_order label. To tackle the phrase-level challenge, we implemented a new model as part of deepQuest (cf. Section 2). The submitted SHEF-ATT-SUM system takes the sum of composing word vectors to create phrase vectors used for regression. This configuration performed the best on the official development set.

The official results are reported in Table 5. While we perform better than the baseline for task 3a, we are not able to beat it at the phrase level. We believe this is because the dataset is too small to train a competitive neural model. There are no other results for gaps prediction.[10]

[10]We did not participate to the source phrases prediction task, since the phrase alignments were not provided by the organisers.

3.4 Task 4: Document-level QE

Task 4 consists in predicting document-level quality scores for MT of product reviews from the Amazon Product Reviews dataset (He and McAuley, 2016). For this task, a selection of Sports and Outdoors product titles and descriptions were machine translated from English into French. The MT system used is a state-of-the-art NMT system. The machine translated documents were annotated with word-level MQM information. The MQM taxonomy has three coarse-grained classes: accuracy, fluency and style. Each error was classified into one of the fine-grained classes within a main class and also according to its severity: minor (it does not change the meaning of the source), major (the meaning was changed by the incorrect word) or critical (besides changing the meaning the error results in a negative effect, e.g. the translation can be seen as offensive).

Document-level scores were devised as follows using the information about the errors and their severities:

$$score = 100 * (1.0 - T_{severity} * \frac{1.0}{N}) \quad (1)$$

where $T_{severity}$ is the sum of the severity weights of all errors in a given document (predefined as minor $= 1.0$, major $= 5.0$ and critical $= 10$) and N is the total number of words in this document.

For training, development and testing, $1,000$, 200 and 269 documents were made available, respectively. The baseline is an SVM model trained with 15 baseline document-level features from QuEst++. Evaluation is done in terms of Pearson's r correlation scores.

Since the MQM scores are at the word level, Equation 1 can also be used to extract scores for sentences. We exploit this feature and create MTL systems trained to predict both sentence and document-level scores. We submitted two systems officially and also report three additional systems. Our systems are listed below, where systems with an * are the official submissions:

- *SHEF-PT (in-domain): `POSTECH` system pre-trained with in-domain data extracted from the English–French part[11] of the Gigaword corpus,[12]
- SHEF-PT (out-domain): `POSTECH` system pre-trained with the Europarl data,
- SHEF-bRNN: our `bRNN` system for document-level QE,
- SHEF-MTL-PT (in-domain): multi-task `POSTECH` pre-trained with the in-domain data, and
- *SHEF-MTL-bRNN: multi-task `bRNN`.

Table 6 shows the evaluation of our systems on the test set in terms of Pearson's r and MAE. The baseline is considerably strong, achieving over 0.5 of correlation and the lowest MAE (56.09). SHEF-PT (in-domain) and SHEF-MTL-PT (in-domain) are the only systems that outperform the baseline. Note that the SHEF-MTL-bRNN system achieved results close to the baseline, even though it does not use any external resources (unlike the SHEF-PT systems and the baseline).

	r	MAE
SHEF-PT (in-domain)	**0.534**	56.23
SHEF-PT (out-domain)	0.511	57.55
SHEF-bRNN	0.468	57.58
SHEF-MTL-PT (in-domain)	0.521	56.60
SHEF-MTL-bRNN	0.473	56.59
Baseline	0.512	**56.09**

Table 6: Evaluation of our systems for task 4 on the test set. We show scores of Pearson's r correlation and MAE.

4 Conclusions

We presented our systems submitted to the WMT18 QE shared task. We experimented with two different architectures: our re-implementation of the `POSTECH` system (SHEF-PT) and our `bRNN` (bi-directional RNNs) approach (SHEF-bRNN). Although SHEF-PT is better than SHEF-bRNN for the majority of the task variants, SHEF-bRNN is still a competitive system and, given its simplicity and independence from external resources, it can be seen as a good alternative for low-resource languages. In addition, it is worth mentioning that SHEF-bRNN requires considerably less training time than SHEF-PT, which may better fit certain scenarios.

Acknowledgments

Carolina Scarton is supported by the EC project SIMPATICO (H2020-EURO-6-2015, grant number 692819). Frédéric Blain is supported by the Amazon Academic Research Awards program.

[11] https://catalog.ldc.upenn.edu/LDC2011T10

[12] ≈300K segments were extracted, using XenC (Rousseau, 2013), as having the best perplexity according to a language model trained on a selection of the English in-domain Amazon reviews (≈200K segments).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of International Conference on Learning Representations (ICLR)*.

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, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 507–517.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. DeepQuest: a framework for neural-based quality estimation. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics: Technical Papers*. The COLING 2017 Organizing Committee.

Nisarg Jhaveri, Manish Gupta, and Vasudeva Varman. 2018. Translation quality estimation for indian languages. In *Proceedings of th 21st International Conference of the European Association for Machine Translation (EAMT)*.

Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017a. Predictor-Estimator: Neural quality estimation based on target word prediction for machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 17(1):3:1–3:22.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017b. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 562–568.

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.

Anthony Rousseau. 2013. XenC: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, (100):73–82.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level Translation Quality Prediction with QuEst++. In *The 53rd Annual Meeting of the Association for Computational Linguistics and Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: System Demonstrations*, Beijing, China.

Ilya Sutskever, Oriol Vinyals, and Quoc V. V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112.

Antonio Toral and Víctor M Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *arXiv preprint arXiv:1701.02901*.