Quality Estimation Shared Task

Findings of the 7^{th} edition

Lucia Specia, Frédéric Blain,
Varvara Logacheva, Ramón F. Estudillo and André Martins

WMT18 – Brussels, Nov 2018

The University Of Sheffield.

Overview

2018 Edition – Goals

- Study the performance of quality estimation approaches on the output of neural MT systems.
- Study the predictability of missing words in the MT output.
- Study the predictability of source words that lead to errors in the MT output.
- Study the effectiveness of manually assigned labels for phrases.
- Study quality predictions for documents from errors annotated at word-level with added severity judgements.

2018 Edition – Tasks

Task 1 HTER prediction at sentence-level
↪ What percentage of the sentence should be post-edited?

Task 2 OK/BAD labelling at word-level (+ gaps, + src words)
↪ Which word(s) in the sentence is/are erroneous?

Task 3 OK/BAD labelling at phrase-level (+ gaps, + src words)
↪ Which phrase(s) in the sentence is/are erroneous?

Task 4 MQM score prediction at document-level
↪ What is the overall quality of the document?

2018 Edition – Participants

ID	Participating team
CMU-LTI	Carnegie Melon University, US [Hu et al., 2018]
JU-USAAR	Jadavpur University, India & Saarland University, 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 [Bicici, 2018]
SHEF	University of Sheffield, UK [Ive et al., 2018]
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]

↪ 10 teams, **111 systems**: up to 2 per team, per subtask & language pair

2018 Edition – Evaluation (NEW)

CodaLab

competitions.codalab.org

- Popular competition hosting platform
- One CODALAB instance per task, sub-tasks as ”phases”
- Continuous evaluation, *immediate* feedback (scoring, ranking)
- Open to new participants, beyond WMT

DATASETS

Datasets – Tasks 1 & 2

Same for sentence- and word-levels: QT21 data [Specia et al., 2017]

Four language pairs, two domains:

- English-German, English-Czech → IT domain
- German-English, English-Latvian → Pharma domain

	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

Datasets – Task 3

Subset of German-English (SMT) data from Task 1

- Translations with HTER=0 and HTER>=.30 are filtered out
- Segmentation into phrases produced by the SMT decoder
- Manually annotated using **BRAT**

Task variant: Task 3a – phrase annotations propagated to word-level

Task 3a	# Sentences	# Words	# BAD
Train.	5,921	126,508	35,532
Dev.	1,000	28,710	6,153
Test	543	7,464	3,089

Task 3b	# Sentences	# Phrases	# BAD
Train.	5,921	50,834	10,451
Dev.	1,000	8,566	1,795
Test	543	4,391	868

Datasets – Task 4

NEW – Product descriptions, from the Amazon Product Review dataset [He and McAuley, 2016, McAuley et al., 2015]

- LP: English–French
- Domain: ”Sports & Outdoors”
- Translations produced by SOTA online NMT system
- Annotated for errors at word-level using Multidimensional Quality Metrics (MQM) taxonomy [Lommel et al., 2014]

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

RESULTS

Task 1 – Sentence-level QE

Task 1 – Sentence-level QE – Settings

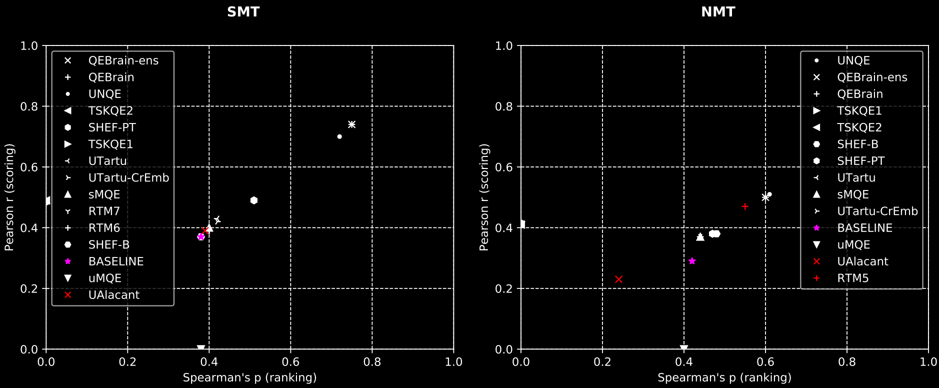
Labels – HTER

Evaluation – Scoring (Pearson's r), Ranking (Spearman's ρ)

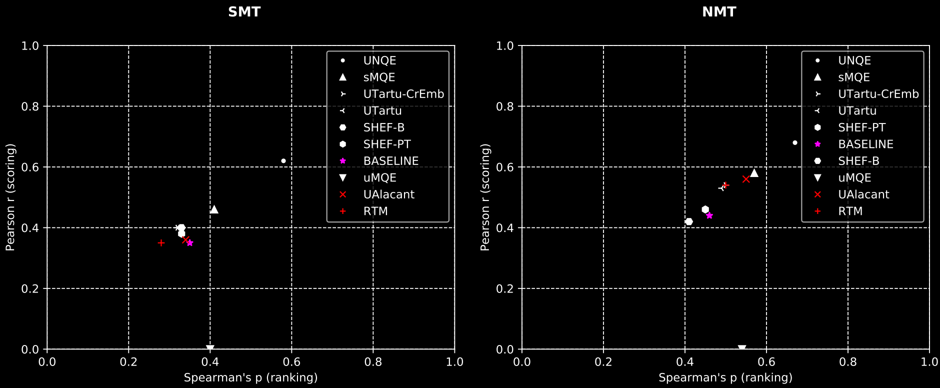
Significance – William's test

BASELINE – QuEst++ for 17 MT system-independent features; SVR with RBF kernel

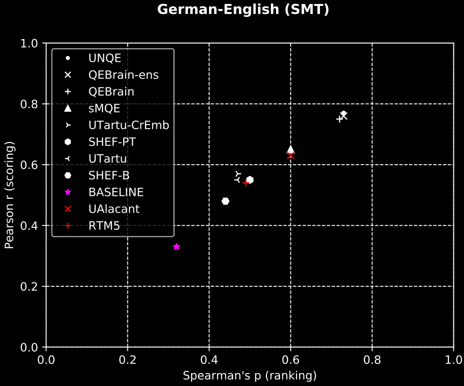
Results – Task 1 – English-German

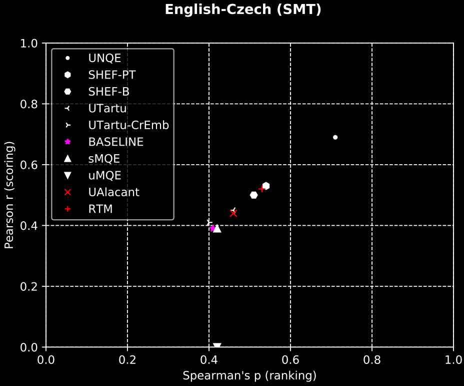


Results – Task 1 – English-Latvian



Results – Task 1 – German-English & English-Czech





Task 1 – Take away message

Task with the most participants (same as previous years)

QEBrain & UNQE systems stand out, winning the task

- QEBrain – conditional LM + Bi-LSTM
  - Multi-head self-attention mechanism and transformer NN to build LM, used as feature extractor
  - Extracted features combined with human-crafted features, and fed into a Bi-LSTM predictive model
  - Greedy ensemble selection method to decrease individual model errors and increase model diversity
- Unified NN architecture for sentence-level QE (UNQE) – Bi-RNN + RNN
  - Bi-RNN with attention mech. – extracts quality vectors
  - RNN – predicts HTER

Interesting margin compared to SHEF-PT (reimplementation of POSTECH, SOTA 2017)

Task 2 – Word-level QE

Task 2 – Word-level QE – Settings

Labels – OK / BAD

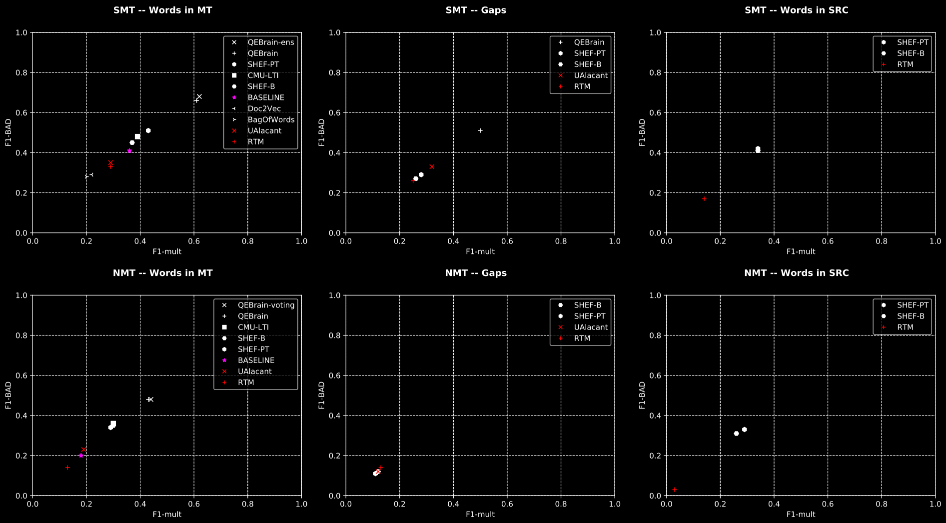
- Target words: OK (=unchanged), BAD (=insertion, substitution)
- Gaps: OK (=genuine gap), BAD (=deletion error(s))
- Source words: OK, BAD (=aligned to substituted or deleted words in target, or missing words)

Evaluation – F_1-OK, F_1-BAD, F_1-mult (=F_1-OK * F_1-BAD)

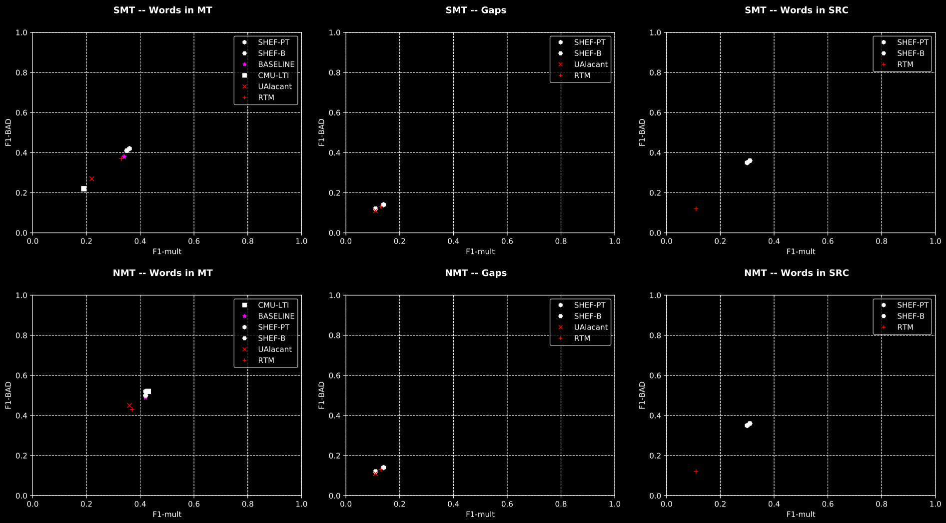
Significance – randomisation test [Yeh, 2000], with Bonferroni correction [Abdi, 2007]

BASELINE – MARMOT with 28 features including language model and context-dependent ones; CRF with passive-aggressive algorithm

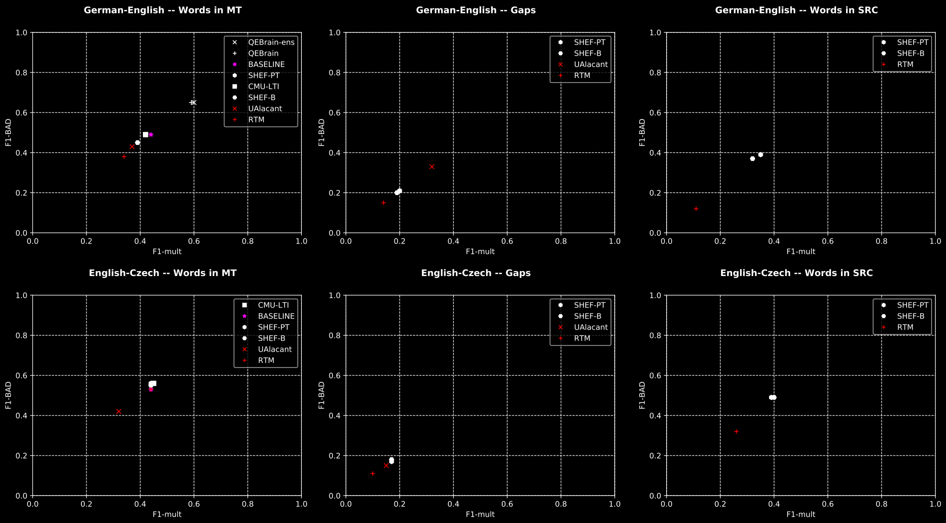
Task 2 – Results – English-German



Task 2 – Results – English-Latvian



Task 2 – Results – German-English & English-Czech



Task 2 – Take away message

Results between tasks 1 & 2 are correlated (continuity from previous years)

English-German & German-English – LPs with the most systems participating
↪ QEBrain system won both subtasks – notable performance for gap error detection (almost double compared to others)
↪ Clear drop in performance from SMT to NMT (English-German)
↪ Low participation to task variants, but correlation with main word-level task

English-Latvian & English-Czech – lower number of participants: due to lower number of resources?

Task 3 – Phrase-level QE

Task 3 – Phrase-level QE – Settings

Labels – OK, BAD, BAD_word_order, BAD_omission

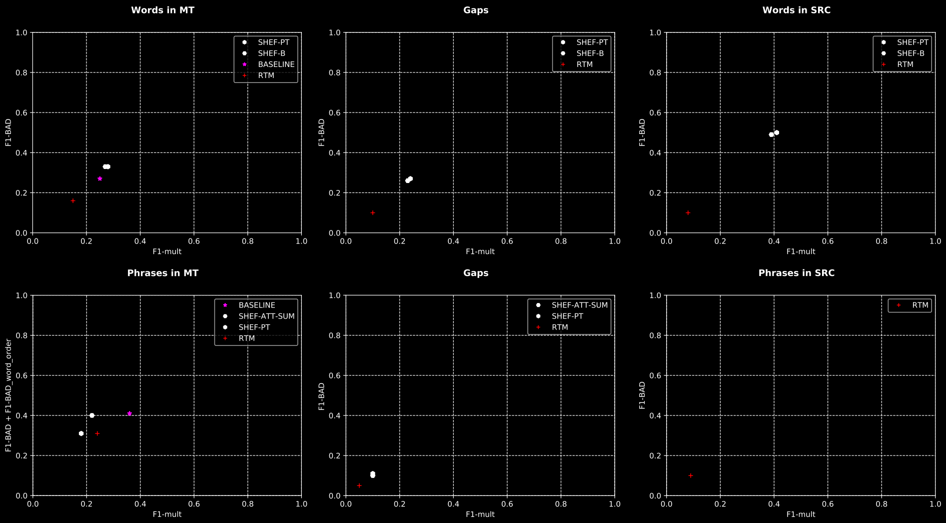
- Target phrases: OK (=unchanged), BAD (=contain one or more errors), BAD_word_order (=in an incorrect position),
- Gaps: OK (=genuine gap), BAD_omission (=missing phrase)
- Source phrases: OK, BAD (=lead to errors in translation)

Evaluation – F_1-OK, F_1-BAD, F_1-mult

Significance – randomisation test with Bonferroni correction, as in Task 2

BASELINE – MARMOT with 72 features adapted from sentence level; CRF with passive-aggressive algorithm

Task 3 – Results – English-German (SMT)



Task 3 – Take away message

Very few submissions (one official + one late)

SHEF-PT & SHEF-ATT-SUM won the task

- SHEF-PT (3a) – Reimplementation of POSTECH system
- SHEF-ATT-SUM (3b) – sum of composing word vectors to create phrase vectors used for regression

Task 3a – general degradation of the F_1-BAD compared to Task 2: word-level from PE *vs.* phrase-level from human

Task 4 – Document-level QE

Task 4 – Document-level QE – Settings

Labels –

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

Evaluation – Pearson's r between the true and predicted document-level scores

BASELINE – QuEst++ for 17 baseline features for document-level, except for the Giza++ related features; SVR with RBF kernel

Task 4 – Results – English-French[1]

Model	Pearson r
● SHEF-PT-indomain	0.53
BASELINE	0.51
SHEF-mtl-bRNN	0.47
RTM_MIX1**	0.11

[1]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.

Task 4 – Take away message

Strong baseline, with high correlation

SHEF-PT-indomain model won the task, outperforming the baseline by a modest margin

- modular architecture wrapping over sentence-level representations from both SHEF-PT & SHEF-B
- SHEF-PT pre-trained with in-domain data selected from the English–French Gigaword corpus

MQM score – document-level score built from word-level annotations: should sentence-level information (*e.g.* importance towards the document) be considered?

DISCUSSION

Performance of QE approaches on the output of neural MT

Task 1 / English-German – More data from SMT than NMT (higher quality, lower HTER) – Top systems & baseline perform better on SMT than NMT – More samples for SMT and/or significant differences in distributions of HTER?

Task 1 / English-Latvian – similar amount of data between SMT and NMT (comparable HTER) – difference between systems is less marked, but trend is inverted: top systems performing better on NMT.

→ QE models seem to be robust to different types of translation, since rankings are the same across datasets.

Discussion

Performance of QE approaches on the output of neural MT

Task 2 – similar trend to Task 1: QE systems for English-German perform better on SMT than on NMT, the inverse is observed for English-Latvian

Task 4 – baseline system performing as well or better than neural-based submissions – First edition, therefore hard to conclude whether the performance of the systems is good enough.

Predictability of missing words in the MT

More difficult than target word error detection, but high scores on SMT data – Unclear on NMT due to too few submissions

Predictability of source words that lead to errors in the MT

Harder problem than detecting errors in the target – Is translation ambiguity responsible?

Discussion

Quality prediction for documents from errors annotated at word-level with added severity judgements

New task and not many systems were submitted – Gap between neural approach and baseline smaller than Task 1 – Would DL architectures tailored for document lead to better results?

2018 Edition – General remarks

- **Largest edition ever organised**
  - ↪ Five LPs, three domains, 111 submitted systems
  - ↪ Various types of annotation (from PE/manual, source/target)
  - ↪ Prediction on neural MT outputs
  - ↪ Prediction on gaps
  - ↪ Prediction on source words
- **Continuous evaluation** (CodaLab)
  - ↪ Future benchmarking on a blind basis

2012-2018 Editions – Lessons learned

- QE task grew in dataset size (2K to 40K)
- QE task diversified in languages (1 to 5)
- QE task covered most granularity levels possible (sentence → sentence, word, phrase, paragraph, document)
- Baselines have been outperformed by most systems by 3-4 years
- Shift from feature-heavy to carefully crafted linguistically motivated features to learned representations
- New challenges with output of neural systems: “adequacy” prediction

Announcement – QE Shared Task 2019

Under new management

QE for **Post-Editing**

$\hookrightarrow$ Predict HTER at sentence-level

$\hookrightarrow$ Predict OK/BAD at word-level

QE for **Diagnostics**

$\hookrightarrow$ Predict MQM erroneous segments, and their error categories

QE for **Scoring**

$\hookrightarrow$ Rank systems as a metric (w/o a reference)

$\hookrightarrow$ Evaluation against human judgements

Thanks.

Feel free to connect with questions[1].

f.blain@sheffield.ac.uk

[1] Poster session today, 11:00–12:30.

APPENDIX

CODALAB – Links to competitions

Task 1 Sentence-level QE
↪ `https://competitions.codalab.org/competitions/19316`

Task 2 Word-level QE
↪ `https://competitions.codalab.org/competitions/19306`

Task 3 Phrase-level QE
↪ `https://competitions.codalab.org/competitions/19308`

Task 4 Document-level QE
↪ `https://competitions.codalab.org/competitions/19309`

CODALAB – How to get the scores for each of my submissions?

After a successful submission, follow those steps:

1. Click on the "Submit / View Results" menu, under the "Participate" tab;
2. Select the subtask you are interested into;
3. For each submission you made, expand its information by clicking on the '+' symbol;
4. Click on "Download output from scoring step", to download the scoring output[1]

[1] unzipped the file corresponding to the submission, and the scores will be into the 'scores.txt' file.

REFERENCES

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

Basu, P., Pal, S., and Naskar, S. K. (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.

Bicici, E. (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.

Duma, M. and Menzel, W. (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.

Etchegoyhen, T., Garcia, E. M., and Azpeitia, A. (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.

He, R. and McAuley, J. (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.

Hu, J., Chang, W.-C., Wu, Y., and Neubig, G. (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.

Ive, J., Scarton, C., Blain, F., and Specia, L. (2018).
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.

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

- McAuley, J., Targett, C., Shi, Q., and Van Den Hengel, A. (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.
- Sánchez-Martíínez, F., Esplà-Gomis, M., and Forcada, M. L. (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.

Specia, L., Harris, K., Blain, F., Burchardt, A., Macketanz, V., Skadina, I., Negri, M., , and Turchi, M. (2017).
Translation quality and productivity: A study on rich morphology languages.
In *Machine Translation Summit XVI*, pages 55–71, Nagoya, Japan.

Wang, J., Fan, K., Li, B., Zhou, F., Chen, B., Shi, Y., and Si, L. (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.

Yankovskaya, E., Tattar, A., and Fishel, M. (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.

Yeh, A. (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.