Quality Estimation Shared Task 2022

Findings of the 11th edition

Chrysoula Zerva[1,2], Frédéric Blain[3], Ricardo Rei[2,4,5], Piyawat Lertvittayakumjorn[6], José G. C. de Souza[4], Steffen Eger[9], Diptesh Kanojia[8], Duarte Alves[2], Constantin Orăsan[8], Marina Fomicheva[7], André F. T. Martins[1,2,7] and Lucia Specia[6,7]

[1]Instituto de Telecomunicações, [2]Instituto Superior Técnico, [3]University of Wolverhampton, [4]Unbabel, [5]INESC-ID, [6]Imperial College London, [7]University of Sheffield [8]University of Surrey, [9]NLLG,Technische Fakultät, Bielefeld University

OVERVIEW

2022 Edition – Highlights

- New **languages covered** in our datasets;
  - English-Marathi (27K segments)
  - English-Yoruba (1K - zero-shot)
- Encourage **language-independent** and even **unsupervised** approaches especially for zero-shot prediction;
- **Fine-grained quality annotation**, informed at word and sentence level using **MQM**: En-De, En-Ru, Zh-En;
- New subtask: **explainable** approaches for Quality Estimation
- Revisited **critical error detection**.

2022 Edition – Tasks

Task 1 Quality estimation at both word- and sentence-level
↪ scoring translations according to their **perceived quality** using direct assessments (DA) and MQM scores as well as binary quality labels on word level.

Task 2 Explainable quality estimation word–level
↪ obtain **word-level rationales** for sentence-level quality scores

Task 3 Critical Error Prediction
↪ binary label at sentence level to indicate whether the sentence **contains one or more critical errors**

2022 Edition – Evaluation

competitions.codalab.org

- One CODALAB instance per sub-task, each language-pair is a different "phase"
- Each participant could submit at most 10 systems for each phase
  - 2 max submissions per day

2022 Edition – Evaluation

competitions.codalab.org

- One CODALAB instance per sub-task, each language-pair is a different "**phase**"
- Each participant could submit at most 10 systems for each phase
  - 2 max submissions per day
- Continuous evaluation, **immediate feedback** (scoring, ranking)
- Open to new participants

2021 Edition – Participants

15 identified teams & 2 anonymous efforts

ID	Affiliations	
Alibaba Translate	DAMO Academy, Alibaba Group & University of Science and Technology of China & CT Lab, University of Macau, China & National University of Singapore, Republic of Singapore	[Bao et al., 2022]
BJTU-Toshiba	Beijing Jiaotong University, China & Toshiba Co., Ltd.	[Huang et al., 2022]
HW-TSC	Huawei Translation Services Center & Nanjing University, China	[Su et al., 2022]
HyperMT - aiXplain	aiXplain	–
IST-Unbabel	INESC-ID & Instituto de Telecomunicações & Instituto Superior Técnico & Unbabel, Portugal	[Rei et al., 2022]
KU X Upstage	Korea University, Korea & Upstage	[Eo et al., 2022]
NJUNLP	Huawei Translation Services Center, China	[Geng et al., 2022]
Papago	Papago, Naver Corp	[Lim and Park, 2022]
UCBerkeley-UMD	University of California, Berkeley & University of Maryland	[Mehandru et al., 2022]
UT-QE	University of Tehran, Iran	[Azadi et al., 2022]
Welocalize-ARC/NKUA	Welocalize Inc, USA & National Kapodistrian University & Athena RC, Greece	[Zafeiridou and Sofianopoulos, 2022]

- **991 submissions** – **Task1**: 81.1%; **Task2**: 16.9%; **Task3**: 2%
- **117 multilingual submissions** – **w/o zero shot**: 65%; **with zero-shot**: 35%

RESULTS & DISCUSSION

Task 1

☞ New setup: 3 subtasks

1. Direct Assessments (DA): continuation of QE setup from previous editions – sentence level
2. Multidimensional Quality Metrics (MQM): new fine-grained annotations – sentence level
3. Word-level: Combined binary OK/BAD word level tags – word level

Task 1 – DA at Sentence-level – Settings

Labels: mean average over z-normalised **Direct Assessments**

Evaluation: Primary scoring: **Spearman's ρ**
Secondary metrics: Pearson's r, MAE, RMSE
Also: Disc footprint, #model parameters, ensemble size – NEW!

Significance: William's test

Baseline: XLM-RoBERTa large Predictor-Estimator approach [Kim et al., 2017]

- implemented in OpenKiwi [Kepler et al., 2019]
- joint learning sentence scores and word quality labels
- fine-tuned language model on train+dev dataset splits

Task 1 DA – Official Results

Model	Multi	Multi w/o En-Yo	En-Cs	En-Ja	En-Mr	Ps-En	Km-En
IST-Unbabel	**0.572**	**0.605**	**0.655**	**0.385**	**0.592**	**0.669**	**0.722**
Papago	0.502	0.571	**0.636**	0.327	**0.604**	0.653	0.671
Alibaba Translate	–	0.585	0.635	0.348	**0.597**	0.657	0.697
Welocalize-ARC/NKUA	0.448	0.506	0.563	0.276	0.444	0.623	–
BASELINE	0.415	0.497	0.560	0.272	0.436	0.579	0.641
lp_sunny‡	0.414	0.485	0.511	0.290	0.395	0.611	0.637
HW-TSC	–	–	0.626	0.341	0.567	0.509	0.661
aiXplain	–	–	0.477	0.274	0.493	–	–
NJUNLP	–	–	–	–	**0.585**	–	–
UCBerkeley-UMD*	–	–	0.285	–	–	–	–

Spearman's ρ: Ranking by average performance for all language pairs

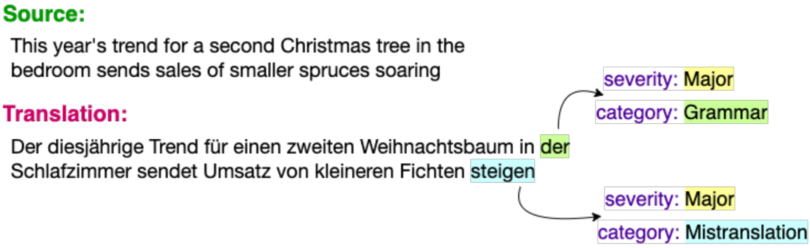
- Best performers: IST-Unbabel & Papago
  - Large pretrained representations + multi-task learning
  - data augmentation/external data + ensembles

☞ Higher performance for into-English translations

Task 1 – MQM at Sentence-level – Example

- Annotations of error spans in-sentence
- Classify by:
  - Severity
  - Category
- Accumulate error penalties according to severity/category for each sentence → final quality score

  ☞ ! score direction is opposite to DA !



Task 1 – MQM at Sentence-level – Settings

Labels inverted and z-normalised MQM scores (to align with DA)

Evaluation Primary scoring: **Spearman's ρ**
Secondary metrics: Pearson's r, MAE, RMSE
Also: Disc footprint, #model parameters, ensemble size – NEW!

Significance William's test

Baseline XLM-RoBERTa large Predictor-Estimator approach [Kim et al., 2017]

- implemented in OpenKiwi [Kepler et al., 2019]
- joint learning sentence scores and word quality labels
- fine-tuned language model on train+dev MQM dataset splits

Task 1 MQM – Official Results

Model	Multi	En-De	En-Ru	Zh-En
IST-Unbabel	**0.474**	0.561	**0.519**	**0.348**
NJUNLP	0.468	**0.635**	**0.474**	**0.296**
Alibaba-Translate	0.456	0.550	**0.505**	**0.347**
Papago	0.449	0.582	**0.496**	**0.325**
lp_sunny ‡	0.415	0.495	0.453	0.298
BASELINE	0.317	0.455	0.333	0.164
BJTU-Toshiba	–	**0.621**	0.434	**0.299**
HW-TSC	–	0.494	0.433	0.369
aiXplain	–	0.376	0.338	0.194
pu_nlp ‡	–	0.611	–	–

Spearman's ρ: Ranking by average performance for all language pairs

- Best performers: IST-Unbabel & NJUNLP & Alibaba & Papago
  - Large pretrained representations + multi-task learning
  - data augmentation/external data + ensembles
- Lower performance compared to DAs

Task 1 – Word Level – Settings

Labels Word-level: **OK / BAD** tag for each target token

- ☞ No SOURCE or GAP tags this year !
- ☞ Aligned tag representations from post-edited and MQM data
- ☞ Convention: attribute deletions to the token on the right.

Evaluation Primary scoring: **Matthews correlation (MCC)**

Secondary metrics: F1-score

Also: Disc footprint, #model parameters, ensemble size– NEW!

Significance Randomization tests + Bonferroni correction

Baseline XLM-RoBERTa large Predictor-Estimator approach [Kim et al., 2017]

- implemented in OpenKiwi [Kepler et al., 2019]
- joint learning sentence scores and word quality labels
- fine-tuned language model on train+dev MQM dataset splits

Task 1 word-level – Official Results

Model	Multi	Multi (w/o En-Yo)	En-Cs	En-Ja	En-Mr	Kh-En	Ps-En	En-De	En-Ru	Zh-En
IST-Unbabel	**0.341**	**0.361**	**0.436**	**0.238**	**0.392**	**0.425**	**0.424**	**0.303**	**0.427**	**0.360**
Papago	0.317	**0.343**	**0.396**	**0.257**	**0.418**	**0.429**	0.374	**0.319**	**0.421**	**0.351**
BASELINE	0.235	0.257	0.325	0.175	0.306	0.402	0.359	0.182	0.203	0.104
HW-TSC	–	0.218	**0.424**	**0.258**	0.351	0.353	0.358	0.274	0.343	0.246
NJUNLP	–	–	–	–	**0.412**	**0.421**	–	**0.352**	**0.390**	**0.308**

Ranking by average performance for all language pairs

- Best performers: IST-Unbabel & Papago & NJUNLP
  - XLM-R large pretrained representations + ensembles
  - Multi-task approaches
  - pseudo-references + external data Metrics Tasks

Task 2 – Explainable QE

Core idea: Translation error identification –> rationale extraction from sentence-level QE systems

*Continues from Explainable Quality Estimation @Eval4NLP 2021

Task 2 – Explainable QE

Core idea: Translation error identification –> rationale extraction from sentence-level QE systems

*Continues from Explainable Quality Estimation @Eval4NLP 2021

- Errors in the input (MT) $\rightarrow$ reasons for imperfect sentence-level scores.
- Each word-level score should signify the contribution of the word to the sentence score reduction

Task 2 – Explainable QE

Core idea: Translation error identification –> rationale extraction from sentence-level QE systems

*Continues from Explainable Quality Estimation @Eval4NLP 2021

- Errors in the input (MT) → reasons for imperfect sentence-level scores.
- Each word-level score should signify the contribution of the word to the sentence score reduction

Requirements:

- ☞ No word-level supervision
- ✓ Sentence level quality score
- ✓ Continuous word-level scores: tokens with the highest scores are expected to correspond to translation errors

Task 2 – Settings

Labels	Word-level: list of continuous scores. Sentence-level: Continuous score (↑)
Evaluation	Primary scoring: **Recall @Top-K** (R-Precision) Secondary metrics: AUC, AP
Significance	Randomisation tests with Bonferroni correction
Baseline	Random word and sentence scores OpenKiwi sentence scores + LIME [Ribeiro et al., 2016]

Task 2 – Official Results

Model	En-Cs	En-Ja	En-Mr	En-Ru	En-De	En-Yo	Km-En	Ps-En	Zh-En
IST-Unbabel	**0.561**	**0.466**	**0.317**	**0.390**	**0.365**	**0.234**	0.665	0.672	**0.379**
HW-TSC	**0.536**	**0.462**	**0.280**	0.313	0.252	–	**0.686**	**0.715**	0.220
BASELINE (OpenKiwi+LIME)	0.417	0.367	0.194	0.135	0.074	0.111	0.580	0.615	0.048
BASELINE (Random)	0.363	0.336	0.167	0.148	0.124	0.144	0.565	0.614	0.093
UT-QE	–	–	–	–	–	–	0.622	0.668	–

Recall@Top-K: Ranking by average performance for all language pairs

- ✻ Best performers: IST-Unbabel & HW-TSC
- ☞ Additional signals:
  - Sparcity of rationales
  - Source-target alignments
- ☞ Correlation between sentence QE performance and explanation performance

Task 3 – Critical error prediction – Description

Core idea: Critical error –> significant deviation from source meaning

Task 3 – Critical error prediction – Description

Core idea: Critical error –> significant deviation from source meaning

☞ We simulated a real-world scenario where <5% of the data has a critical error with one of the following categories:

- **Additions**: Deviation where only partially supported by the source.
- **Deletions**: Deviation where part of the source sentence is ignored.
- **Named Entities**: Deviation in named entities.
- **Meaning**: Deviation in sentence meaning (e.g. introduction or removal of a negation)
- **Numbers**: Deviation in units (number/date/time or unit).

Task 3 – Critical error prediction – Settings

Settings	unconstrained | constrained (training)
Labels	Binary: ERR | NOT
Evaluation	Primary scoring: Matthews Correlation (MCC)
	Secondary metrics: F1-score
	Also: Disc footprint, #model parameters, ensemble – NEW!
Significance	William's test
Baseline	COMET-QE (constrained)[1]
	XLM-RoBERTa classifier (unconstrained)

[1] *wmt21-comet-qe-da*

Task 3 – Official Results

Model	En-De (Cons)	En-De (Uncons)	Pt-En (Cons)	Pt-En (Uncons)
KU X Upstage	–	**0.964**	–	**0.984**
IST-Unbabel	**0.564**	–	**0.721**	–
BASELINE	0.074	0.855	-0.001	0.934
aiXplain	–	0.219	–	0.179

MCC: Ranking by **average** performance for all language pairs

- Best performers: **KU X Upstage**
- Constrained setting more challenging – realistic?
- Revise setup for future editions?

2022 Edition – General remarks

Promising results: moderate to strong correlation for QE subtasks

2022 Edition – General remarks

Promising results: moderate to strong correlation for QE subtasks

- **Overall:** Multi-task, multi-lingual systems ++ unsupervised but resource heavy: ensembles of large models
- ☞ How to deal with the trade-off between performance and size?

2022 Edition – General remarks

Promising results: moderate to strong correlation for QE subtasks

- **Overall:** Multi-task, multi-lingual systems ++ unsupervised but resource heavy: ensembles of large models
- ☞ How to deal with the trade-off between performance and size?

- **MQM:** Fine-grained annotations seem promising
- ☞ Revise score aggregations?
- ☞ Consider evaluating correlation with human judgements together with robustness to critical errors

2022 Edition – General remarks

Promising results: moderate to strong correlation for QE subtasks

- ✻ **Overall:** Multi-task, multi-lingual systems ++ unsupervised but resource heavy: ensembles of large models
- ☞ How to deal with the trade-off between performance and size?

- ✻ **MQM:** Fine-grained annotations seem promising
- ☞ Revise score aggregations?
- ☞ Consider evaluating correlation with human judgements together with robustness to critical errors

- ✻ **Zero-shot**: More challenging setup ("surprise language")
- ☞ more language pairs?
- ☞ how to mitigate restrictions from pre-trained models?

2022 Edition – General remarks

Promising results: moderate to strong correlation for QE subtasks

- ✻ **Overall:** Multi-task, multi-lingual systems ++ unsupervised but resource heavy: ensembles of large models
- ☞ How to deal with the trade-off between performance and size?

- ✻ **MQM:** Fine-grained annotations seem promising
- ☞ Revise score aggregations?
- ☞ Consider evaluating correlation with human judgements together with robustness to critical errors

- ✻ **Zero-shot**: More challenging setup ("surprise language")
- ☞ more language pairs?
- ☞ how to mitigate restrictions from pre-trained models?

- ✻ **Explainability**: Promising results but challenging to come up with a representative setup
- ☞ Improve evaluation and baseline scheme?

ALL the results, gold labels, submissions and baseline predictions are freely available!

https://wmt-qe-task.github.io/

Stay tuned for the 12th edition!

Thank you!

Feel free to connect during the Q&A session.

chrysoula.zerva@tecnico.ulisboa.pt

REFERENCES

Azadi, F., Faili, H., and Dousti, M. J. (2022).
Mismatching-Aware Unsupervised Translation Quality Estimation for Low-Resource Languages.
arXiv preprint arXiv:2208.00463.

Bao, K., Wan, Y., Liu, D., Yang, B., Lei, W., He, X., Wong, D. F., and Xie, J. (2022).
Alibaba-translate china's submission for wmt 2022 quality estimation shared task.
In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Eo, S., Park, C., Moon, H., Seo, J., and Lim, H. (2022).
KU X Upstage's submission for the WMT22 Quality Estimation: Critical Error Detection Shared Task.
In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Geng, X., Zhang, Y., Huang, S., Tao, S., Yang, H., and Chen, J. (2022).
NJUNLP's Participation for the WMT2022 Quality Estimation Shared Task.
In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Huang, H., Di, H., Li, C., Wu, H., Oushi, K., Chen, Y., Liu, J., and Xu, J. (2022).
BJTU-Toshiba's Submission to WMT22 Quality Estimation Shared Task.
In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Kepler, F., Trénous, J., Treviso, M., Vera, M., and Martins, A. F. T. (2019).
OpenKiwi: An open source framework for quality estimation.
In *Proceedings of ACL 2019 System Demonstrations*.

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

Lim, S. S. and Park, J. (2022).
Papago's submission to the wmt22 quality estimation shared task.
In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Mehandru, N., Carpuat, M., and Selehi, N. (2022).
Quality Estimation by Backtranslation at the WMT 2022 Quality Estimation Task.
In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., de Souza, J. G. C., Glushkova, T., Alves, D., Lavie, A., Coheur, L., and Martins, A. F. T. (2022).
CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task.
In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016).
" why should i trust you?" explaining the predictions of any classifier.
In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Su, C., Ma, M., Tao, S., Yang, H., Zhang, M., Geng, X., Huang, S., Guo, J., Wang, M., and Li, Y. (2022).
CrossQE: HW-TSC 2022 Submission for the Quality Estimation Shared.
In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Zafeiridou, E. and Sofianopoulos, S. (2022).
Welocalize-ARC/NKUA's Submission to the WMT 2022 Quality Estimation Shared Task.
In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

2022 Edition – Data breakdown

Language Pairs	Sentences Train / Dev / Test22	Tokens Train / Dev / Test22	DA	PE	MQM	CE	Data Source
En-De [1]	8,000 / 1,000 / –	131,499 / 16,545 / –	✓	✓			Wikipedia
En-Zh	8,000 / 1,000 / –	131,892 / 16,637 / –	✓	✓			Wikipedia
Ru-En	8,000 / 1,000 / –	94,221 / 11,650 / –	✓	✓			Reddit
Ro-En	8,000 / 1,000 / –	137,466 / 17,359 / –	✓	✓			Wikipedia
Et-En	8,000 / 1,000 / –	112,503 / 14,044 / –	✓	✓			Wikipedia
Ne-En	8,000 / 1,000 / –	120,078 / 15,017 / –	✓	✓			Wikipedia
Si-En	8,000 / 1,000 / –	125,223 / 15,709 / –	✓	✓			Wikipedia
En-Mr	26,000 / 1,000 / 1,000	690,532 / 27,049 / 26,253	✓	✓			
Ps-En	– / 1,000 / 1,000	– / 27,045 / 27,414	✓	✓			Wikipedia
Km-En	– / 1,000 / 1,000	– / 21,981 / 22,048	✓	✓			Wikipedia
En-Ja	– / 1,000 / 1,000	– / 20,626 / 20,646	✓	✓			Wikipedia
En-Cs	– / 1,000 / 1,000	– / 20,394 /20,244	✓	✓			Wikipedia
En-Yo	– / – / 1,010	– / – / 21,238	✓	✓			
En-De [2]	28,909 / 1,005 / 511	839,473 / 24,373 / 13,220			✓		WMT-newstest
En-Ru	15,628 / 1,005 / 511	357,452 / 24,373 / 13,220			✓		WMT-newstest
Zh-En	35,327 / 1,019 / 505	1,586,883 / 51,969 / 15,602			✓		WMT-newstest
En-De	155,511 / 17,280 / 500	8,193,693 / 915,061 / 27,771				✓	News-Commentary
Pt-En	39,926 / 4,437 / 500	2,281,515 / 253,594 / 29,794				✓	News-Commentary

Statistics of the data used for Task 1 (DA), Task 2 (PE) and Task 3 (CE)

☞ NEW! test sets for Task 1 (DA): English-Marathi and English-Yoruba

2022 Edition – Data breakdown

Language Pairs	Sentences Train / Dev / Test22	Tokens Train / Dev / Test22	DA	PE	MQM	CE	Data Source
En-De [1]	8,000 / 1,000 / –	131,499 / 16,545 / –	✓	✓			Wikipedia
En-Zh	8,000 / 1,000 / –	131,892 / 16,637 / –	✓	✓			Wikipedia
Ru-En	8,000 / 1,000 / –	94,221 / 11,650 / –	✓	✓			Reddit
Ro-En	8,000 / 1,000 / –	137,466 / 17,359 / –	✓	✓			Wikipedia
Et-En	8,000 / 1,000 / –	112,503 / 14,044 / –	✓	✓			Wikipedia
Ne-En	8,000 / 1,000 / –	120,078 / 15,017 / –	✓	✓			Wikipedia
Si-En	8,000 / 1,000 / –	125,223 / 15,709 / –	✓	✓			Wikipedia
En-Mr	26,000 / 1,000 / 1,000	690,532 / 27,049 / 26,253	✓	✓			
Ps-En	– / 1,000 / 1,000	– / 27,045 / 27,414	✓	✓			Wikipedia
Km-En	– / 1,000 / 1,000	– / 21,981 / 22,048	✓	✓			Wikipedia
En-Ja	– / 1,000 / 1,000	– / 20,626 / 20,646	✓	✓			Wikipedia
En-Cs	– / 1,000 / 1,000	– / 20,394 /20,244	✓	✓			Wikipedia
En-Yo	– / – / 1,010	– / – / 21,238	✓	✓			
En-De [2]	28,909 / 1,005 / 511	839,473 / 24,373 / 13,220			✓		WMT-newstest
En-Ru	15,628 / 1,005 / 511	357,452 / 24,373 / 13,220			✓		WMT-newstest
Zh-En	35,327 / 1,019 / 505	1,586,883 / 51,969 / 15,602			✓		WMT-newstest
En-De	155,511 / 17,280 / 500	8,193,693 / 915,061 / 27,771				✓	News-Commentary
Pt-En	39,926 / 4,437 / 500	2,281,515 / 253,594 / 29,794				✓	News-Commentary

Statistics of the data used for Task 1 (DA), Task 2 (PE) and Task 3 (CE)

☞ NEW! test sets for Task 1 (DA): English-Marathi and English-Yoruba

Task 1 – Word Level – Settings

Labels Word-level: **OK / BAD** tag for each word

- ☞ No SOURCE or GAP tags this year !

Task 1 – Word Level – Settings

Labels Word-level: **OK / BAD** tag for each word

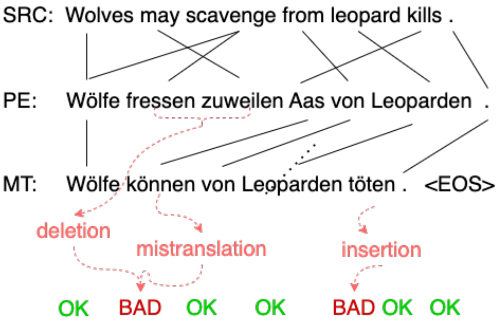
- ☞ No SOURCE or GAP tags this year !
- ☞ Aligned tag representations from post-edited and MQM data
- ☞ Convention: attribute deletions to the token on the right.

Task 1 – Word Level – Settings

Labels Word-level: **OK / BAD** tag for each word

- No SOURCE or GAP tags this year !
- Aligned tag representations from post-edited and MQM data
- Convention: attribute deletions to the token on the right.

Post-edit example:

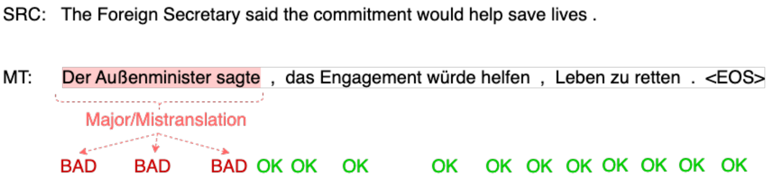


Task 1 – Word Level – Settings

Labels Word-level: **OK / BAD** tag for each word

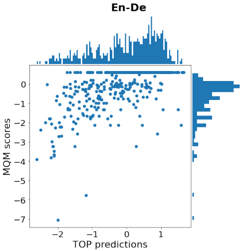
- ☞ No SOURCE or GAP tags this year !
- ☞ Aligned tag representations from post-edited and MQM data
- ☞ Convention: attribute deletions to the token on the right.

MQM example:

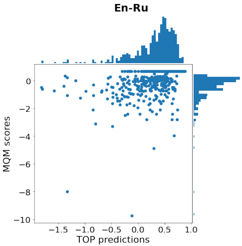


Task 1 – MQM vs DA analysis

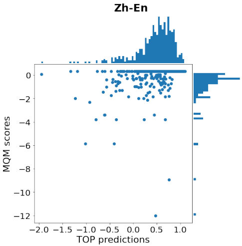
MQM languages



(a) En-De

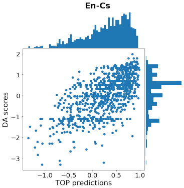


(b) En-Ru

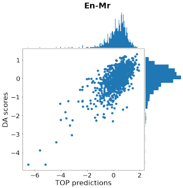


(c) Zh-En

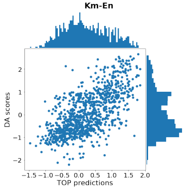
DA/PE languages



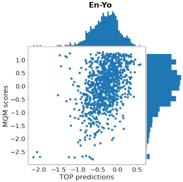
(d) En-Cs



(e) En-Mr



(f) Km-En



(g) En-Yo