Article:

Fomicheva, M., Sun, S., Fonseca, E.R. et al. (6 more authors) (Submitted: 2020) MLQE-PE : a multilingual quality estimation and post-editing dataset. arXiv. (Submitted)

MLQE-PE: A Multilingual Quality Estimation and Post-Editing Dataset

Marina Fomicheva,[1*] Shuo Sun,[2*] Erick Fonseca,[3] Frédéric Blain,[1,4]
Vishrav Chaudhary,[5] Francisco Guzmán,[5] Nina Lopatina,[6] Lucia Specia,[1,7] André F. T. Martins[3,8]
[1]University of Sheffield, [2]Johns Hopkins University,
[3]Instituto de Telecomunicações, [4]University of Wolverhampton,
[5]Facebook AI, [6]IQT Labs, [7]Imperial College London, [8]Unbabel
{m.fomicheva, f.blain, l.specia}@sheffield.ac.uk, ssun32@jhu.edu,
erick.fonseca@lx.it.pt, {fguzman, vishrav}@fb.com,
nlopatina@iqt.org, andre.martins@unbabel.com

Abstract

We present MLQE-PE, a new dataset for Machine Translation (MT) Quality Estimation (QE) and Automatic Post-Editing (APE). The dataset contains seven language pairs, with human labels for 9,000 translations per language pair in the following formats: sentence-level direct assessments and post-editing effort, and word-level good/bad labels. It also contains the post-edited sentences, as well as titles of the articles where the sentences were extracted from, and the neural MT models used to translate the text.

1 Introduction

Translation quality estimation (QE) is the task of evaluating a translation system's quality without access to reference translations (Blatz et al., 2004; Specia et al., 2018b). This task has numerous applications: deciding if a sentence or document that has been automatically translated is ready to be sent to the final user or if it needs to be post-edited by a human, flagging passages with potentially critical mistakes, using it as a metric for translation quality when a human reference is not available, or in the context of computer-aided translation interfaces, highlighting text that needs human revision and estimating the human effort.

Due to its high relevance, QE has been the subject of evaluation campaigns in the Conference for Machine Translation (WMT) since 2014 (Bojar et al., 2014; Specia et al., 2018a; Fonseca et al., 2019), where datasets in various language pairs have been created containing source sentences, their automatic translations, and human post-edited text. However, the currently existing data has several shortcomings. First, the MT system used to produce the translations is not publicly available, which makes it impossible to develop the so-called glass-box approaches to QE and exploit model confidence (or conversely, uncertainty) of the MT system or look into its internal states. Second, the quality assessments have been either produced based on the difference between the MT output and the post-edited text (e.g., through the human translation error rate metric, HTER, or by marking individual words with OK or BAD labels), or by direct human assessments, but not both—which raises the question of how much these two quality assessments correlate. Third, most datasets have focused exclusively on high-resource language pairs, where it is often the case that many sentences are correctly translated; however, medium and low-resource settings are the ones where QE would be particularly useful, since it is where MT currently presents serious challenges. Finally, most of these datasets focus on a specific domain, such as IT or life sciences, where translations are generated by a domain-specific MT model, which also tends to result in most sentences being translated with high-quality.

To overcome the limitations stated above, we introduce MLQE-PE, the first multilingual quality estimation and post-editing dataset that combines the following features:

- It includes access to the state-of-the-art neural MT (NMT) models built with an open-source toolkit (`fairseq`, Ott et al. (2019)), that were used to produce the translations in the dataset. This opens the door to uncertainty-based and glass-box approaches to QE.

- It combines both direct assessments of MT quality and post-edits. This allows combining two sorts of quality assessments: how good a translation is and how much effort is necessary to correct it. Moreover, the post-edited sentences can be used for training and evalu-

[*] Equal contribution.

ating automatic post-editing systems, another important task considered in WMT campaigns (Chatterjee et al., 2019).

- It contains the titles of the Wikipedia articles where the original sentences were extracted from, thus allowing to take document-level context into account when predicting sentence-level or word-level MT quality.

- It includes 7 language pairs, mixing high-resource language pairs (English-German – En-De and English-Chinese – En-Zh, and Russian-English – Ru-En), medium-resource (Romanian-English – Ro-En, and Estonian-English – Et-En) and low-resource ones (Nepali-English – Ne-En and Sinhala-English – Si-En).

This dataset was created with contributions from different institutions: Facebook, University of Sheffield and Imperial College on the selection of the Wikipedia articles and sentences, building of NMT models, preparation and annotation of sentences with direct assessments for 6 of the language pairs (En-De, En-Zh, Ro-En, Et-En, Ne-En, Si-En), as well as the creation of reference translations for Et-En; IQT Labs on the equivalent effort for Ru-En; Unbabel and Instituto de Telecomunicações on the annotation of En-De and En-Zh sentences with post-editing. The current version of MLQE-PE is publicly available from `https://github.com/sheffieldnlp/mlqe-pe`.

2 Data Collection and Statistics

We briefly describe the data collection and preparation process. Table 1 presents some statistics about the MLQE-PE dataset.

Data collection. For the most part, the dataset is derived from Wikipedia articles (with exception of Russian-English, described below). The source sentences were collected from Wikipedia articles following the sampling process outlined in FLORES (Guzmán et al., 2019). First, we sampled documents from Wikipedia for English, Estonian, Romanian, Sinhala, and Nepali. Second, we selected the top 100 documents containing the largest number of sentences that are: (i) in the intended source language according to a language-id classifier[1] and (ii) have the length between 50 and 150 characters. In addition, we filtered out sentences that have been released as part of recent Wikipedia parallel corpora (Schwenk et al., 2019), ensuring that our dataset is not part of parallel data commonly used for NMT training.

For every language, we randomly selected 10K sentences[2] from the sampled documents and then translated them into English. For translation we used the NMT models trained based on the standard Transformer architecture (Vaswani et al., 2017) with `fairseq` toolkit (Ott et al., 2019) (see Fomicheva et al. (2020b) for a detailed description of the models). For German and Chinese we selected 20K sentences from the top 100 documents in English Wikipedia. To ensure sufficient representation of high- and low-quality translations for these high-resource language pairs, we selected the sentences with minimal lexical overlap with respect to the NMT training data. Specifically, we extracted content words for each sentence in the data used for training the NMT models and in the Wikipedia data. We computed perplexity scores for the Wikipedia sentences given the NMT training data. Finally, we sampled 20K from available Wikipedia sentences weighted by the perplexity scores.

In addition, we collected human reference translations for a 1K subset of Estonian-English dev/test data. Two reference translations were generated independently by two professional translators. This part of the dataset allows for comparing reference-free MT evaluation with reference-based approaches (see Fomicheva et al. (2020a) for details).

The Russian-English data collection followed a slightly different set up collected by collaborators from IQT Labs.[3] The original sentences were collected from multiple sources in order to gather a varied sample of data in different domains that are still challenging for current NMT systems. Data sources include: Russian proverbs and Reddit data from various subreddits, particularly those focused on topics of politics and religion. We included Reddit data since colloquial text is a challenge for MT. We included Russian proverbs from WikiQuotes to test MT on short sentences with unconventional grammar. We used the Reddit API and queried the most recent 1000 posts at the time, and the

[1] `https://fasttext.cc`

[2] 1K of these sentences will be kept as blind test set and released later.

[3] We note that Facebook was not involved in the collection of the Russian-English data.

Languages	Sentences	Tokens	DA	PE
En-De	7,000/1,000/1,000	114,980 / 16,519 / 16,371	✓	✓
En-Zh	7,000/1,000/1,000	115,585 / 16,307 / 16,765	✓	✓
Ru-En	7,000/1,000/1,000	82,229 / 11,992 / 11,760	✓	
Ro-En	7,000/1,000/1,000	120,198 / 17,268 / 17,001	✓	
Et-En	7,000/1,000/1,000	98,080 / 14,423 / 14,358	✓	
Ne-En	7,000/1,000/1,000	104,934 / 15,144 / 14,770	✓	
Si-En	7,000/1,000/1,000	109,515 / 15,708 / 15,821	✓	

Table 1: Statistics of the MLQE-PE dataset. Number of sentences and tokens are shown for train/development/test partitions, respectively. Number of tokens refer to the source side.

	En-De	En-Zh
Train	-0.131	-0.291
Dev	-0.344	-0.251
Test	-0.291	-0.214

Table 2: Pearson correlation between direct assessments and HTER scores for the test partition of the dataset

most recent 1000 comments in each of the selected subreddits. We automatically split the posts into sentences and then reviewed these manually. Markdown was removed and HTML unencoded. We removed sentences shorter than 15 characters or longer than 500 characters. We also removed sentences that did not have a source link. Table 3 shows the number of segments corresponding to each data source and the corresponding average direct assessment score.

NMT models Transformer-based (Vaswani et al., 2017) NMT models were trained for all languages using the `fairseq` toolkit.[4] For **Et-En**, **Ro-En**, **En-De** and **En-Zh** we trained the MT models based on the standard Transformer architecture following the implementation details described in Ott et al. (2018). We used publicly available MT datasets such as Paracrawl (Esplà et al., 2019) and Europarl (Koehn, 2005). For **Ru-En**, translations were produced with the already existing Transformer-based NMT model described in Ng et al. (2019) and available at https://github.com/pytorch/fairseq/tree/master/examples/wmt19. **Si-En** and **Ne-En** MT systems were trained based on Big-Transformer architecture as defined in Vaswani et al. (2017). For these low-resource language pairs, the models were trained following the FLORES semi-supervised setting (Guzmán et al., 2019)[5], which involves two iterations of backtranslation using the source and the target monolingual data. The data used for training the NMT models is available from http://www.statmt.org/wmt20/quality-estimation-task.html. We provide access to the information from the NMT model used to generate the translations: model score for the sentence and log probabilities for words, as well as the NMT systems themselves.

Direct assessments. To collect human quality judgments, we followed the FLORES setup (Guzmán et al., 2019) inspired by the work of Graham et al. (2013). Specifically, the annotators were asked to rate translation quality for each sentence on a 0–100 scale, where the 0–10 range represents an incorrect translation; 11–29, a translation that contains a few correct keywords, but the overall meaning is different from the source; 30–50, a translation with major mistakes; 51–69, a translation which is understandable and conveys the overall meaning of the source but contains typos or grammatical errors; 70–90, a translation that closely preserves the semantics of the source sentence; and 91–100, a perfect translation.

Each segment was evaluated independently by three professional translators from a single language service provider. To improve annotation consistency, any evaluation in which the range of scores among the raters was above 30 points was rejected, and an additional rater was requested to replace the most diverging translation rating until convergence was achieved. To further increase the reliability of the test and development partitions of the dataset, we requested an additional set of three

[4] https://github.com/pytorch/fairseq

[5] https://github.com/facebookresearch/flores/blob/master/reproduce.sh

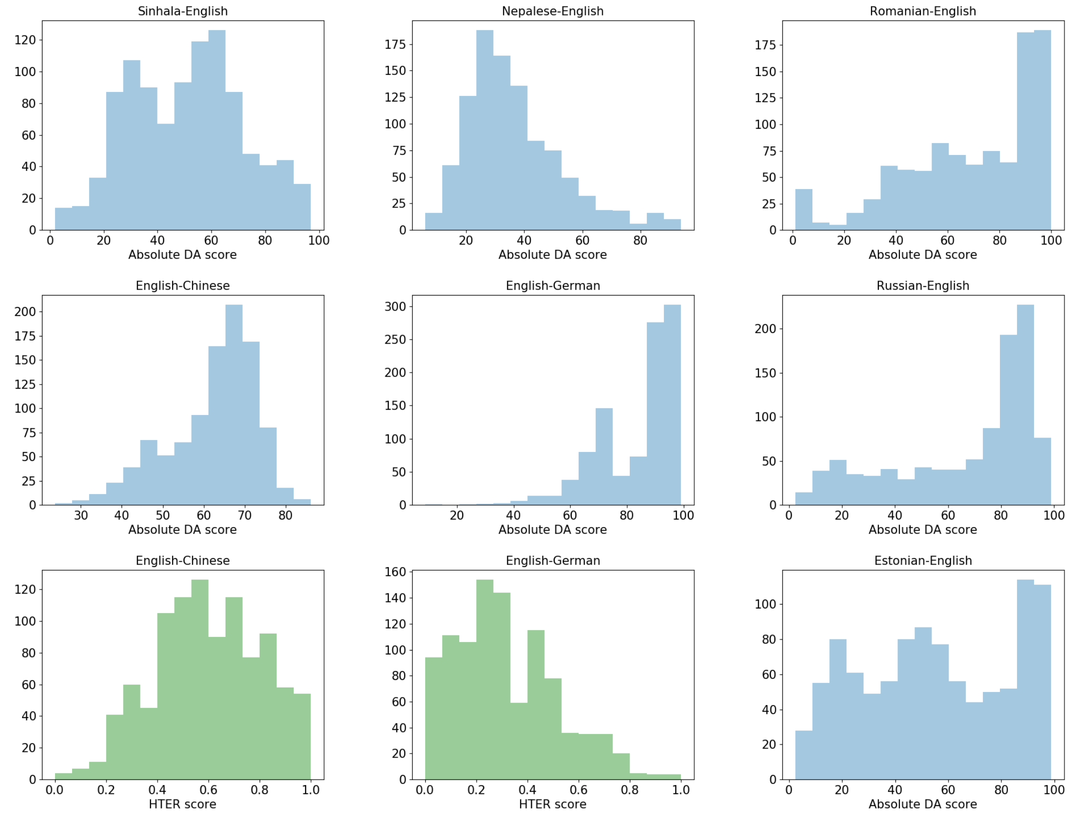

Figure 1: Distribution of direct assessment and HTER scores on the test partition of the dataset

	Count	DA
www.reddit.com/r/antireligious	2,155	75.6
www.reddit.com/r/PikabuPolitics	1,753	77.7
www.reddit.com/r/rupolitika	1,422	80.1
www.reddit.com/r/ru	2,171	74.0
wikiquote.org/wiki	2,499	41.1

Table 3: Number of sentences and average absolute direct assessment (DA) score for each data source in the Ru-En dataset

annotations from a different group of annotators (i.e., from another language service provider) following the same annotation protocol, thus resulting in a total of six annotations per segment.

Raw human scores were converted into z-scores, that is, standardized according to each individual annotator's overall mean and standard deviation. The scores collected for each segment were averaged to obtain the final score. Such setting allows for the fact that annotators may genuinely disagree on some aspects of quality.

Human post-editing. For the two high-resource language pairs (En-De and En-Zh), the translated sentences have been post-edited by two human translators, paid editors from the Unbabel community. The two human translators had no access to the direct assessments above. The average human translation error rate between the machine translated text and the post-edited text was 0.32 for En-De and 0.62 for En-Zh.

Figure 1 shows the distribution of direct assessments and HTER scores for the test partition of the dataset. First, we note that the distribution of direct assessment scores is very different across language pairs. This illustrates the variety of the collected data in terms of MT output quality. For low-resource language pairs there are more sentences with low direct assessment scores, whereas in the case of high-resource language pairs the vast majority of translations received a high score. In

Type	Text	Scores
Source	He wakes up in a cage, and enjoys rubbing the rusted bars.	
MT	他在笼子里醒来, 喜欢擦生锈的酒吧.	DA = 33
PE	他在笼子里醒来, 喜欢摩擦生锈的铁条。	HTER = 0.33
MT gloss	He wakes up in a cage, and enjoys rubbing the rusted **pub**.	
PE gloss	He wakes up in a cage, and enjoys rubbing the rusted **metal bar**.	

Table 4: Example of the discrepancy between HTER and DA annotation tasks: low DA score (low quality) but low HTER score (minimal post-editing).

Type	Text	Scores
Source	The two battled to a standstill and eventually rendered one another comatose.	
MT	这两个人的战斗陷入停顿, 最后彼此昏迷不已.	DA = 73
PE	两人对战陷入僵局，最后双双昏倒。	HTER = 1.00
MT gloss	The two people's battle fell into a standstill, finally both were in a coma.	
PE gloss	The two people battled to a standstill and both fell into a coma.	

Table 5: Example of the discrepancy between HTER and DA annotation tasks: high DA score (high quality) but high HTER score (substantial post-editing).

particular, En-De has a very peaked distribution with very little variability in quality.

Second, direct assessments and HTER for the same data behave quite differently. The distribution of HTER scores for En-De is also skewed towards the high-quality end (lower HTER scores indicate better quality) but is much smoother, meaning that some sentences that received high scores during direct assessment evaluation were still corrected by the post-editors.

To gain a better understanding of this difference, Table 2 shows Pearson correlation between direct assessments and HTER scores. Figure 2 contains the corresponding scatter plots for the test partition of the dataset. We observe that the correlation is fairly low for both language pairs.

Direct quality assessment and post-editing give two different perspectives on MT quality. Table 4 shows an example where direct assessment and HTER lead to a different interpretation of quality. Direct assessment score is low as the MT output contains a serious error that distorts the meaning of the sentence: "bars" (as in "metal bars") is translated as "pub". However the sentence is easy to post-edit as the error involves only one word to be replaced, resulting in a low HTER score. Table 5 illustrates the opposite: MT output was assigned a high direct assessment score, but the HTER score is also high, indicating that substantial changes were introduced during post-editing. The post-edited version is more fluent, whereas the MT output is a more literal rendering of the source sentence, but the meaning is preserved and, therefore, it received a high direct assessment score.

Word-level labels In the datasets containing post-edit annotation, we also obtained word-level labels for fine-grained post-editing effort estimation. Both the source and MT sides have them.

In order to generate them, we first align source and MT outputs using FastAlign[6], trained on the same data as the NMT models, and compute the shortest edit distances between MT and post-edited texts with Tercom[7]; this effectively informs us which words were deleted, inserted or replaced. Then, any word w_s in the source aligned to a word w_m in MT that was kept in the post-edit receives a tag OK; if w_s is not aligned with any other word in MT or if w_m was deleted in the post-edit, it is tagged BAD. Thus, BAD tags in the source side indicate which words caused MT errors.

For the MT side, we tag both words and the gaps between them, indicating whether a missing additional word should have been there. Any word w_m aligned to another word w_p in the post-edit receives a tag OK; words deleted or replaced are tagged BAD. Any gap g between words in the MT output, before the first word or after the last one receives a tag OK if no word w_p is inserted in there, and BAD otherwise.

Statistics for word-level tags are shown in Table 6. We see that most sentences in the dataset have at least one BAD tag; in the case of En-Zh, it is nearly all of them. The overall amount of BAD tags is also higher in the En-Zh data, especially in the source side.

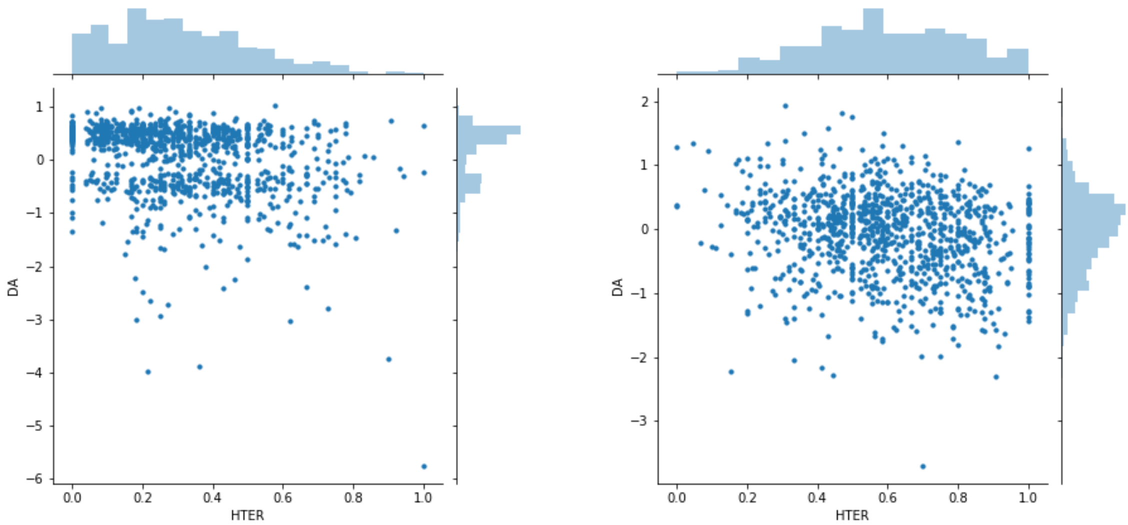

Figure 2: Correlation between HTER and DA scores for the test partition of the dataset for English-German (left) and English-Chinese (right)

		Source		Target	
		BAD tags	Sentences	BAD tags	Sentences
En-De	Train	26.95%	92.27%	16.02%	93.60%
	Dev	25.79%	91.90%	15.49%	93.40%
	Test	25.77%	92.60%	15.53%	93.60%
En-Zh	Train	53.59%	99.71%	30.53%	99.81%
	Dev	50.92%	99.50%	28.98%	99.70%
	Test	49.99%	99.50%	28.85%	99.70%

Table 6: Ratio of BAD tags in the word-level data for the different splits of the dataset (third and fifth columns), and ratio of sentences containing at least one such tag (fourth and sixth columns).

	Words in MT			**Words in SRC**		
Languages	MCC	F_1-BAD	F_1-OK	MCC	F_1-BAD	F_1-OK
En-De	0.358	0.468	0.879	0.266	0.477	0.779
En-Zh	0.509	0.658	0.849	0.270	0.682	0.547

Table 7: Performance at **word-level** of Predictor-Estimator baseline models for each label and language pair of the MLQE-PE dataset.

3 Baseline performance

We report the performance of baseline systems trained on the MLQE-PE data. We used the Predictor-Estimator approach (Kim et al., 2017), implemented in OpenKiwi toolkit (Kepler et al., 2019). The Predictor model was trained on the same parallel data as the NMT systems for each language pair (made availalbe by the WMT20 Shared Task on Quality Estimation)[8], while the Estimator model was trained on the 7,000 QE labelled data for each task. Both models were trained using default configurations and parameters. Tables 7 and 8 present the performance of our baseline systems for each label and language pair, for word- and sentence-level predictions respectively.

[6] https://github.com/clab/fast_align
[7] http://www.cs.umd.edu/~snover/tercom/

[8] http://www.statmt.org/wmt20/quality-estimation-task.html

Languages	Pearson r	MAE	RMSE
Direct Assessment			
En-De	0.146	0.679	0.967
En-Zh	0.190	0.885	1.068
Ru-En	0.548	0.825	1.193
Ro-En	0.685	0.760	1.052
Et-En	0.477	0.918	1.138
Ne-En	0.386	0.735	0.871
Si-En	0.374	0.752	0.898
HTER			
En-De	0.392	0.150	0.190
En-Zh	0.506	0.147	0.181

Table 8: Performance at **sentence-level** of Predictor-Estimator baseline models for each label and language pair of the MLQE-PE dataset.

4 Conclusions

We introduced MLQE-PE, a new dataset that was mainly created to be used for the tasks of quality estimation (sentence and word-level prediction) and automatic post-editing. It contains data in seven language pairs, direct assessment and post-editing-based sentence-level labels, as well as binary good/bad word-level labels. In addition, a subset of the data contains independently created reference translations, which can be used, for example, for machine translation evaluation. The dataset is freely available and was already used for the WMT2020 shared tasks on Quality Estimation and Automatic Post-Editing.[9]

We hope that this data will foster further work on these and other tasks, such as uncertainty estimation and model calibration. We also hope it will sparkle interest from researchers who may want to contribute related resources, i.e., more data, different languages, etc.

Acknowledgments

Marina Fomicheva, Frédéric Blain and Lucia Specia were supported by funding from the Bergamot project (EU H2020 Grant No. 825303). André Martins and Erick Fonseca were funded by the P2020 programs Unbabel4EU (contract 042671) and MAIA (contract 045909), by the European Research Council (ERC StG DeepSPIN 758969), and by the Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020. We would like to thank Camila Pohlmann and the Unbabel community team for monitoring the post-editing process. We also thank Mark Fishel from the University of Tartu for providing the Estonian reference translations.

[9] http://www.statmt.org/wmt20/ape-task.html

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *Proc. of the International Conference on Computational Linguistics*, page 315.

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 *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the wmt 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Marina Fomicheva, Lucia Specia, and Francisco Guzmán. 2020a. Multi-hypothesis machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1218–1232.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020b. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computation Linguistics*.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6097–6110, Hong Kong, China. Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. 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 *Conference on Machine Translation (WMT)*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair wmt19 news translation task submission. In *Proc. of WMT*, pages 1–4.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André FT Martins. 2018a. Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018b. Quality Estimation for Machine Translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.