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

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, André F. T. Martins and Lucia Specia

LREC, 2022
We present a multilingual dataset with sentence and word level annotations for Machine Translation (MT) Quality Estimation (QE)
We present a **multilingual** dataset with sentence and word level annotations for Machine Translation (MT) Quality Estimation (QE)
TL;DR

We present a multilingual dataset with sentence and word level annotations for Machine Translation (MT) Quality Estimation (QE)
We present a multilingual dataset with sentence and word level annotations for Machine Translation (MT) Quality Estimation (QE)

Wolves may scavenge from leopard kills

Wolves may kill from leopards

Wölfe können von Leoparden töten
We present a multilingual dataset with sentence and word level annotations for Machine Translation (MT) Quality Estimation (QE)

Wolves may scavenge from leopard kills

Wolves may kill from leopards

Wölfe können von Leoparden töten

Corrected: Wölfe fressen zuweilen Aas von Leoparden

Quality score: 0.45
Why Quality Estimation?
Why Quality Estimation?

Wolves may scavenge from leopard kills  
source sentence

Wölfe können von Leoparden töten  
MT sentence

Wolves may kill from leopards
Why Quality Estimation?

Wolves may scavenge from leopard kills

**source sentence**

Wölfe können von Leoparden töten

**MT sentence**

Wolves may kill from leopards

Wölfe fressen zuweilen Aas von Leoparden

**human reference**

Wolves sometimes eat carrion of leopards
Why Quality Estimation?

Wolves may scavenge from leopard kills

Wölfe können von Leoparden töten

source sentence

Wolves may kill from leopards

MT sentence

Wölfe fressen zuweilen Aas von Leoparden

Wolves sometimes eat carrion of leopards

human reference

MT Evaluation

BLEU

ChRF

BERTScore

COMET

BLEURT
Why Quality Estimation?

Wolves may scavenge from leopard kills

source sentence

Wölfe können von Leoparden töten

MT sentence

Wolves may kill from leopards

Wölfe fressen zuweilen Aas von Leoparden

Wolves sometimes eat carrion of leopards

human reference

MT Evaluation

Allow for quality estimation methods without the need for human references

BLEU

ChRF

BERTScore

COMET

BLEURT
Why Quality Estimation?

MT Evaluation is not always possible/desired:
Why Quality Estimation?

MT Evaluation is not always possible/desired:

- Expensive and time consuming to obtain references
Why Quality Estimation?

MT Evaluation is not always possible/desired:

- Expensive and time consuming to obtain references
- On-the-fly translation
  - Flag potentially critical errors
  - Decide which segments need human-editing
Why Quality Estimation?

MT Evaluation is not always possible/desired:

▶ Expensive and time consuming to obtain references
▶ On-the-fly translation
  ▶ Flag potentially critical errors
  ▶ Decide which segments need human-editing
▶ Zero-shot applications:
  ▶ Apply to low-resource languages
  ▶ Adapt to domains without human references
MLQE-PE Quality Estimation Data Annotations

For each segment (source - MT sentence pair) we provide:
MLQE-PE Quality Estimation Data Annotations

For each segment (source - MT sentence pair) we provide:

**Sentence level scores:**

- Sentence level direct assessments (DA scores)
  - 3 annotators per segment
  - Mean z-score as final value
- Human-targeted Translation Edit Rate (HTER)
  - Minimum number of edits needed to reach from the MT to the post-edited sentence
  - Normalised by sentence length
    - 0-1 scale
MLQE-PE Quality Estimation Data Annotations

For each segment (source - MT sentence pair) we provide:

**Sentence level scores:**
- Sentence level direct assessments (DA scores)
  - 3 annotators per segment → mean z-score as final value
  - Scale 0-100:

![Score scale with categories](Image)

- incorrect translation
- wrong meaning - few correct keywords
- major mistakes
- understandable but grammar errors/typos
- correct semantics - minor errors
- perfect translation
MLQE-PE Quality Estimation Data Annotations

For each segment (source - MT sentence pair) we provide:

**Sentence level scores:**

- **Sentence level direct assessments (DA scores)**
  - 3 annotators per segment → mean z-score as final value
  - Scale 0-100:

    ![Sentence level direct assessments](image)

- **Human-targeted Translation Edit Rate (HTER)**
  - Minimum number of edits needed to reach from the MT to the post-edited sentence
  - Normalised by sentence length → 0-1 scale
MLQE-PE Quality Estimation Data Annotations

For each segment (source - MT sentence pair) we provide:

Sentence level scores:
- Sentence level direct assessments (DA scores)
  - 3 annotators per segment $\rightarrow$ mean z-score as final value
  - Scale 0-100:

- Human-targeted Translation Edit Rate (HTER)
  - Minimum number of edits needed to reach from the MT to the post-edited sentence
  - Normalised by sentence length $\rightarrow$ 0-1 scale

Word level scores:
MLQE-PE Quality Estimation Data Annotations

For each segment (source - MT sentence pair) we provide:

**Sentence level scores:**

- **Sentence level direct assessments (DA scores)**
  - 3 annotators per segment → mean z-score as final value
  - Scale 0-100:

- **Human-targeted Translation Edit Rate (HTER)**
  - Minimum number of edits needed to reach from the MT to the post-edited sentence
  - Normalised by sentence length → 0-1 scale

**Word level scores:**

- **Binary OK or BAD tags**
  - On the target (MT) tokens: Wrong or irrelevant tokens
  - On the target gaps (between tokens): Missing tokens
  - On the source tokens: Mistranslated or non-translated tokens
Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?
Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?
- Extract alignments between PE and MT, SRC
  - MT-PE $\rightarrow$ Monolingual: TERcom
  - Source-MT
  - Source-PE $\}$ Bilingual: SimAlign

Wolves may scavenge from leopard kills.

Fomicheva, Sun, Fonseca, Zerva, Blain, Chaudhary, Guzmán, Lopatina, Ma
Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?

- Extract alignments between PE and MT, SRC
  - MT-PE \(\rightarrow\) Monolingual: TERcom
  - Source-MT
  - Source-PE

\{\}
Bilingual: SimAlign
Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?

- Extract alignments between PE and MT, SRC
  - MT-PE \rightarrow \text{Monolingual: TERcom}
  - Source-MT
  - Source-PE \{ \text{Bilingual: SimAlign} \}

Wolves may scavenge from leopard kills.

Wölfe fressen zuweilen Aas von Leoparden.

\{ Fomicheva, Sun, Fonseca, Zerva, Blain, Chaudhary, Guzmán, Lopatina, Ma }
Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?
- Extract alignments between PE and MT, SRC
  - MT-PE → Monolingual: TERcom
  - Source-MT
  - Source-PE
  - Bilingual: SimAlign

Wolves may scavenge from leopard kills .

SRC
Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?

- Extract alignments between PE and MT, SRC
  - MT-PE → Monolingual: TERcom
  - Source-MT
  - Source-PE
  \} Bilingual: SimAlign

Wolves may scavenge from leopard kills.

SRC

Wölfe können von Leoparden töten.

MT
Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?

- Extract alignments between PE and MT, SRC
  - MT-PE $\rightarrow$ Monolingual: TERcom
  - Source-MT
  - Source-PE $\{\}$ Bilingual: SimAlign

Wolves may scavenge from leopard kills.

Wölfe fressen zuweilen Aas von Leoparden.

Wölfe können von Leoparden töten.
Introduction

Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?

- Extract alignments between PE and MT, SRC
  - MT-PE → Monolingual: TERcom
  - Source-MT
  - Source-PE

Bilingual: SimAlign

Wolves may scavenge from leopard kills.

Wölfe fressen zuweilen Aas von Leoparden.

Wölfe können von Leoparden töten.

Fomicheva, Sun, Fonseca, Zerva, Blain, Chaudhary, Guzmán, Lopatina, Ma
Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?

- Extract alignments between PE and MT, SRC
  - MT-PE $\rightarrow$ Monolingual: TERcom
  - Source-MT
  - Source-PE $\rightarrow$ Bilingual: SimAlign

---

Wolves may scavenge from leopard kills.

\[
\begin{align*}
Wölfe & \rightarrow \text{OK} \\
Wölfen & \rightarrow \text{BAD} \\
\text{fressen} & \rightarrow \text{OK} \\
\text{zuweilen} & \rightarrow \text{OK} \\
\text{Aas} & \rightarrow \text{OK} \\
\text{von} & \rightarrow \text{BAD} \\
\text{Leoparden} & \rightarrow \text{OK} \\
\text{töten} & \rightarrow \text{BAD} \\
\end{align*}
\]
Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?

- Extract alignments between PE and MT, SRC
  - MT-PE → Monolingual: TERcom
  - Source-MT
  - Source-PE → Bilingual: SimAlign

Wolves may scavenge from leopard kills

Wölfe fressen zuweilen Aas von Leoparden

Wölfe können von Leoparden töten

OK BAD OK OK BAD OK OK

mistranslation
Introduction

Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?

- Extract alignments between PE and MT, SRC
  - MT-PE \rightarrow \text{Monolingual: TERcom}
  - Source-MT
  - Source-PE \quad \{ \text{Bilingual: SimAlign} \}

Wolves may scavenge from leopard kills.

\begin{align*}
\text{SRC} & \quad \text{PE} \quad \text{MT} \\
\text{Wölfe} & \quad \text{fressen} \quad \text{zuweilen} \quad \text{Aas} \quad \text{von} \quad \text{Leoparden} \\
\text{Wolves} & \quad \text{may} \quad \text{scavenge} \quad \text{from} \quad \text{leopard} \quad \text{kills} \\
\text{\ }
\end{align*}

\begin{align*}
\text{OK} & \quad \text{BAD} \quad \text{OK} \quad \text{OK} \quad \text{BAD} \quad \text{OK} \quad \text{OK} \quad \text{OK} \quad \text{OK} \quad \text{OK} \\
\text{OK} & \quad \text{BAD} \quad \text{OK} \quad \text{OK} \quad \text{OK} \quad \text{OK} \quad \text{OK} \quad \text{OK} \quad \text{OK} \quad \text{OK} \\
\text{deletion} & \quad \text{mistranslation}
\end{align*}
Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?

- Extract alignments between PE and MT, SRC
  - MT-PE $\rightarrow$ Monolingual: TERcom
  - Source-MT
  - Source-PE

\[\begin{align*}
\text{Wolves} & \quad \text{may} & \quad \text{scavenge} & \quad \text{from} & \quad \text{leopard} & \quad \text{kills} \\
\text{Wölfe} & \quad \text{fressen} & \quad \text{zuweilen} & \quad \text{Aas} & \quad \text{von} & \quad \text{Leoparden} \\
\text{Wölfe} & \quad \text{können} & \quad \text{von} & \quad \text{Leoparden} & \quad \text{töten} & \\
\end{align*}\]

- OK
- BAD
- deletion
- mistranslation
- insertion
Word-level tag extraction

How do we extract OK and BAD tags from post-edited sentences?

- Extract alignments between PE and MT, SRC
  - MT-PE $\rightarrow$ Monolingual: TERcom
  - Source-MT
  - Source-PE $\rightarrow$ Bilingual: SimAlign

```
Wolves may scavenge from leopard kills.
```

```
Wölfe fressen zuweilen Aas von Leoparden.
```

```
OK BAD BAD OK OK BAD OK
Wolves may scavenge from leopard kills.
```

```
Wölfe fressen zuweilen Aas von Leoparden.
```

```
OK BAD OK OK OK OK OK
```

OK BAD OK OK OK OK OK

decomposition

deletion

mistranslation

insertion
Annotations by language pair
Annotations by language pair

- English - German
- English - Chinese
- English - Japanese
- Russian - English
- Estonian - English
- English - Czech
- Romanian - English
- Pashto - English
- Sinhala - English
- Khmer - English
- Nepali - English

High resource

Low resource

no resource
Annotations by language pair

Sentence Distribution per LP

High resource
- English - German
- English - Chinese
- English - Japanese
- Russian - English
- Estonian - English
- English - Czech
- Romanian - English
- Pashto - English
- Sinhala - English
- Khmer - English
- Nepali - English

Low Resource

no resource
Additional information
Additional information

- NMT models used to obtain the translations
  - Enable glass-box approaches
  - NMT uncertainty as a proxy to quality
Additional information

- NMT models used to obtain the translations
  - Enable glass-box approaches
  - NMT uncertainty as a proxy to quality
- Independent annotator scores (DA) for each segment
  - Annotator disagreement $\rightarrow$ STD scores
  - Aleatoric uncertainty
  - Proxy to noisy/complex segments
Introduction

Additional information

- NMT models used to obtain the translations
  - Enable glass-box approaches
  - NMT uncertainty as a proxy to quality
- Independent annotator scores (DA) for each segment
  - Annotator disagreement $\rightarrow$ STD scores
  - Aleatoric uncertainty
  - Proxy to noisy/complex segments

\[ \text{motivate uncertainty aware approaches} \]
Additional information

• NMT models used to obtain the translations
  • Enable glass-box approaches
  • NMT uncertainty as a proxy to quality
• Independent annotator scores (DA) for each segment
  • Annotator disagreement $\rightarrow$ STD scores
  • Aleatoric uncertainty
  • Proxy to noisy/complex segments

• Document level information
  • Provision of document ids
  • Use title/surrounding sentences
Additional information

• NMT models used to obtain the translations
  • Enable glass-box approaches
  • NMT uncertainty as a proxy to quality
• Independent annotator scores (DA) for each segment
  • Annotator disagreement $\rightarrow$ STD scores
  • Aleatoric uncertainty
  • Proxy to noisy/complex segments

• Document level information
  • Provision of document ids
  • Use title/surrounding sentences

motivate
uncertainty
aware
approaches

motivate
context
aware
approaches
Additional information

- NMT models used to obtain the translations
  - Enable glass-box approaches
  - NMT uncertainty as a proxy to quality
- Independent annotator scores (DA) for each segment
  - Annotator disagreement $\rightarrow$ STD scores
  - Aleatoric uncertainty
  - Proxy to noisy/complex segments

- Document level information
  - Provision of document ids
  - Use title/surrounding sentences

✓ Useful features for quality estimation

motivate uncertainty aware approaches

motivate context aware approaches
DA-HTER score correlations

High-resource language pairs:

- Different score distributions
- HTER scores (horiz.) are skewed to zero
- Upper-left corner → high-quality translations
DA-HTER score correlations

Mid & Low resource language pairs:

- En-Zh
- Ne-En
- Km-En
- Et-En
- Si-En
- Pe-En
## More on scores and correlations

<table>
<thead>
<tr>
<th></th>
<th>Avg. DA</th>
<th>Avg. HTER</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-De</td>
<td>82.61</td>
<td>0.18</td>
<td>-0.42</td>
<td>-0.48</td>
</tr>
<tr>
<td>Ro-En</td>
<td>69.18</td>
<td>0.24</td>
<td>-0.76</td>
<td>-0.71</td>
</tr>
<tr>
<td>En-Ja</td>
<td>67.96</td>
<td>0.36</td>
<td>-0.14</td>
<td>-0.11</td>
</tr>
<tr>
<td>En-Cs</td>
<td>66.94</td>
<td>0.26</td>
<td>-0.41</td>
<td>-0.46</td>
</tr>
<tr>
<td>En-Zh</td>
<td>62.86</td>
<td>0.23</td>
<td>-0.21</td>
<td>-0.16</td>
</tr>
<tr>
<td>Et-En</td>
<td>60.09</td>
<td>0.29</td>
<td>-0.61</td>
<td>-0.63</td>
</tr>
<tr>
<td>Ps-En</td>
<td>53.53</td>
<td>0.53</td>
<td>-0.71</td>
<td>-0.67</td>
</tr>
<tr>
<td>Si-En</td>
<td>51.42</td>
<td>0.59</td>
<td>-0.29</td>
<td>-0.28</td>
</tr>
<tr>
<td>Km-En</td>
<td>46.58</td>
<td>0.65</td>
<td>-0.49</td>
<td>-0.43</td>
</tr>
<tr>
<td>Ne-En</td>
<td>36.51</td>
<td>0.66</td>
<td>-0.54</td>
<td>-0.49</td>
</tr>
<tr>
<td>Ru-En</td>
<td>68.67</td>
<td>0.23</td>
<td>-0.51</td>
<td>-0.47</td>
</tr>
</tbody>
</table>
Discrepancies

Case 1: Minimal post-editing

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.

He wakes up in a cage, and enjoys rubbing the rusted pub.

PE

他在笼子里醒来, 喜欢摩擦生锈的铁条。

He wakes up in a cage, and enjoys rubbing the rusted metal bar.

- Average DA score: 0.33 → low quality
- HTER score from PE: 0.33 → high quality
Introduction

Discrepancies

Case 2: Heavy post-editing

The two battled to a standstill and eventually rendered one another comatose.

The two people’s battle fell into a standstill, finally both were in a coma.

Average DA score: 0.73 → high quality

HTER score from PE: 1.00 → low quality
Baseline model

One of the main goals is to support the development of better QE models

- Predictor-Estimator architecture
  - Based on OpenKiwi
- Single head for sentence level DA
- Multitasking for:
  - Sentence-level HTER
  - Word-level tags
## Baseline sentence-level results

<table>
<thead>
<tr>
<th>Languages</th>
<th>Pearson $r \uparrow$</th>
<th>RMSE $\downarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direct Assessment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>En-De</td>
<td>0.403</td>
<td>0.433</td>
</tr>
<tr>
<td>En-Zh</td>
<td>0.525</td>
<td>0.534</td>
</tr>
<tr>
<td>Ru-En</td>
<td>0.677</td>
<td>0.492</td>
</tr>
<tr>
<td>Ro-En</td>
<td>0.818</td>
<td>0.408</td>
</tr>
<tr>
<td>Et-En</td>
<td>0.660</td>
<td>0.543</td>
</tr>
<tr>
<td>Ne-En</td>
<td>0.738</td>
<td>0.524</td>
</tr>
<tr>
<td>Si-En</td>
<td>0.513</td>
<td>0.626</td>
</tr>
<tr>
<td>En-Cs</td>
<td>0.352</td>
<td>0.686</td>
</tr>
<tr>
<td>En-Ja</td>
<td>0.230</td>
<td>0.617</td>
</tr>
<tr>
<td>Km-En</td>
<td>0.562</td>
<td>0.614</td>
</tr>
<tr>
<td>Ps-En</td>
<td>0.476</td>
<td>0.711</td>
</tr>
<tr>
<td><strong>AVG</strong></td>
<td>0.541</td>
<td>0.562</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Languages</th>
<th>Pearson $r \uparrow$</th>
<th>RMSE $\downarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HTER</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>En-De</td>
<td>0.529</td>
<td>0.129</td>
</tr>
<tr>
<td>En-Zh</td>
<td>0.282</td>
<td>0.246</td>
</tr>
<tr>
<td>Ru-En</td>
<td>0.448</td>
<td>0.188</td>
</tr>
<tr>
<td>Ro-En</td>
<td>0.862</td>
<td>0.111</td>
</tr>
<tr>
<td>Et-En</td>
<td>0.714</td>
<td>0.149</td>
</tr>
<tr>
<td>Ne-En</td>
<td>0.626</td>
<td>0.160</td>
</tr>
<tr>
<td>Si-En</td>
<td>0.607</td>
<td>0.159</td>
</tr>
<tr>
<td>En-Cs</td>
<td>0.306</td>
<td>0.206</td>
</tr>
<tr>
<td>En-Ja</td>
<td>0.098</td>
<td>0.232</td>
</tr>
<tr>
<td>Km-En</td>
<td>0.576</td>
<td>0.196</td>
</tr>
<tr>
<td>Ps-En</td>
<td>0.503</td>
<td>0.290</td>
</tr>
<tr>
<td><strong>AVG</strong></td>
<td>0.502</td>
<td>0.188</td>
</tr>
</tbody>
</table>
## Baseline word-level results

<table>
<thead>
<tr>
<th>Languages</th>
<th>Words in MT</th>
<th></th>
<th>Words in SRC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MCC↑ F$_1$-Multi↑</td>
<td></td>
<td>MCC↑ F$_1$-Multi↑</td>
<td></td>
</tr>
<tr>
<td>En-De</td>
<td>0.370 0.415</td>
<td></td>
<td>0.322 0.363</td>
<td></td>
</tr>
<tr>
<td>En-Zh</td>
<td>0.247 0.308</td>
<td></td>
<td>0.241 0.295</td>
<td></td>
</tr>
<tr>
<td>Ru-En</td>
<td>0.256 0.319</td>
<td></td>
<td>0.251 0.292</td>
<td></td>
</tr>
<tr>
<td>Ro-En</td>
<td>0.536 0.553</td>
<td></td>
<td>0.511 0.539</td>
<td></td>
</tr>
<tr>
<td>Et-En</td>
<td>0.461 0.512</td>
<td></td>
<td>0.405 0.459</td>
<td></td>
</tr>
<tr>
<td>Ne-En</td>
<td>0.440 0.483</td>
<td></td>
<td>0.390 0.438</td>
<td></td>
</tr>
<tr>
<td>Si-En</td>
<td>0.425 0.456</td>
<td></td>
<td>0.335 0.379</td>
<td></td>
</tr>
<tr>
<td>En-Cs</td>
<td>0.273 0.372</td>
<td></td>
<td>0.224 0.312</td>
<td></td>
</tr>
<tr>
<td>En-Ja</td>
<td>0.131 0.217</td>
<td></td>
<td>0.175 0.272</td>
<td></td>
</tr>
<tr>
<td>Km-En</td>
<td>0.351 0.409</td>
<td></td>
<td>0.279 0.355</td>
<td></td>
</tr>
<tr>
<td>Ps-En</td>
<td>0.313 0.425</td>
<td></td>
<td>0.249 0.361</td>
<td></td>
</tr>
<tr>
<td><strong>AVG</strong></td>
<td><strong>0.346 0.402</strong></td>
<td></td>
<td><strong>0.307 0.370</strong></td>
<td></td>
</tr>
</tbody>
</table>
In practice

Already used:

✓ WMT Quality Estimation Shared Tasks
  • 2020 Edition: En-De, En-Zh, Et-En, Ne-En, Ro-En, Ru-En, Si-En
  • 2021 Edition: En-Cs, En-Ja, Km-En, Ps-En

✓ WMT Automated Post Editing Shared Tasks
  • 2020 Edition: En-De, En-Zh
  • 2021 Edition: En-De, En-Zh

✓ Eval4NLP Explainable Quality Estimation Task
  • 2021 Edition: Et-En, Ro-En

Maybe also:

▶ Catastrophic error detection
▶ Active learning approaches
▶ Context-aware quality estimation
In practice

Already used:

✓ WMT Quality Estimation Shared Tasks
  • 2020 Edition: En-De, En-Zh, Et-En, Ne-En, Ro-En, Ru-En, Si-En
  • 2021 Edition: En-De, En-Zh, Et-En, Ne-En, Ro-En, Ru-En, Si-En, En-Cs, En-Ja, Km-En, Ps-En
In practice

Already used:

✓ WMT Quality Estimation Shared Tasks
  • 2020 Edition: En-De, En-Zh, Et-En, Ne-En, Ro-En, Ru-En, Si-En
  • 2021 Edition: En-De, En-Zh, Et-En, Ne-En, Ro-En, Ru-En, Si-En, En-Cs, En-Ja, Km-En, Ps-En

✓ WMT Automated Post Editing Shared Tasks
  • 2020 Edition: En-De, En-Zh
  • 2021 Edition: En-De, En-Zh
In practice

Already used:

✓ WMT Quality Estimation Shared Tasks
  • 2020 Edition: En-De, En-Zh, Et-En, Ne-En, Ro-En, Ru-En, Si-En
  • 2021 Edition: En-De, En-Zh, Et-En, Ne-En, Ro-En, Ru-En, Si-En, En-Cs, En-Ja, Km-En, Ps-En

✓ WMT Automated Post Editing Shared Tasks
  • 2020 Edition: En-De, En-Zh
  • 2021 Edition: En-De, En-Zh

✓ Eval4NLP Explainable Quality Estimation Task
  • 2021 Edition: Et-En, Ro-En
In practice

Already used:

✓ WMT Quality Estimation Shared Tasks
  • 2020 Edition: En-De, En-Zh, Et-En, Ne-En, Ro-En, Ru-En, Si-En
  • 2021 Edition: En-De, En-Zh, Et-En, Ne-En, Ro-En, Ru-En, Si-En, En-Cs, En-Ja, Km-En, Ps-En

✓ WMT Automated Post Editing Shared Tasks
  • 2020 Edition: En-De, En-Zh
  • 2021 Edition: En-De, En-Zh

✓ Eval4NLP Explainable Quality Estimation Task
  • 2021 Edition: Et-En, Ro-En

Maybe also:

▶ Catastrophic error detection
▶ Active learning approaches
▶ Context-aware quality estimation
MLQE-PE is intended to be a continuously expanding resource
That’s not all

MLQE-PE is intended to be a continuously expanding resource

✓ Contribute resources:
  • New language pairs (especially low-resource)
  • New domains - challenge sets
  • Additional annotations - references

✓ Use
  • New tasks
  • Compare performance on existing tasks

✓ Provide feedback :)

Fomicheva, Sun, Fonseca, Zerva, Blain, Chaudhary, Guzmán, Lopatina, Ma

MLQE-PE : A Multilingual QE and PE Dataset
Thank you!

THANK YOU!

Questions?

[QR Code for arxiv paper]

[QR Code for github code]