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

TL;DR

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

TL;DR

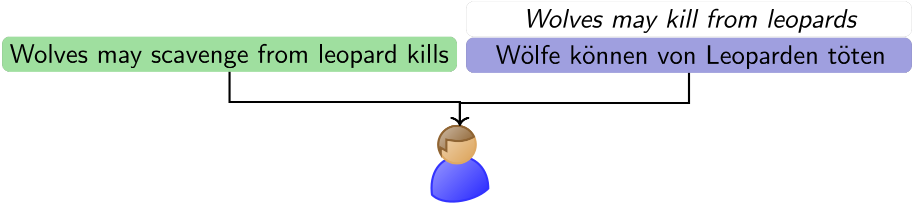
We present a multilingual dataset with sentence and word level annotations for Machine Translation (MT) Quality Estimation (QE)

TL;DR

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

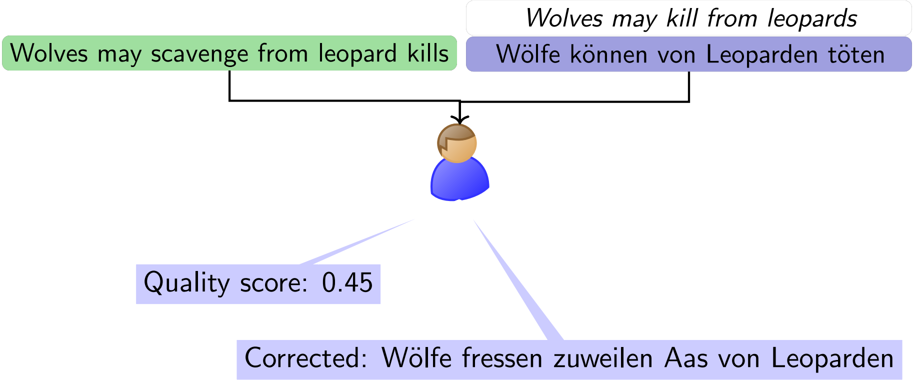
TL;DR

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



TL;DR

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



Why Quality Estimation?

Why Quality Estimation?

Wolves may scavenge from leopard kills

source sentence

Wölfe können von Leoparden töten

Wolves may kill from leopards

MT sentence

Why Quality Estimation?

Wolves may scavenge from leopard kills

source sentence

Wölfe können von Leoparden töten

Wolves may kill from leopards

MT sentence

Wölfe fressen zuweilen Aas von Leoparden

Wolves sometimes eat carrion of leopards

human reference

Why Quality Estimation?

Wolves may scavenge from leopard kills

source sentence

Wölfe können von Leoparden töten

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 BLEURT COMET

Why Quality Estimation?

Wolves may scavenge from leopard kills

Wölfe können von Leoparden töten

Wolves may kill from leopards

source sentence

MT sentence

Quality estimation

Allow for quality estimation methods **without** the need for human references

Wölfe fressen zuweilen Aas von Leoparden

Wolves sometimes eat carrion of leopards

human reference

MT Evaluation

BLEU **ChRF** **BERTScore** **BLEURT** **COMET**

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:

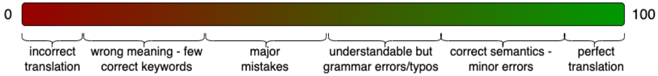
Sentence level scores:

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 $\longrightarrow$ mean z-score as final value
  - Scale 0-100:

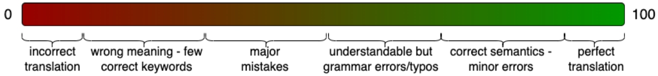


MLQE-PE Quality Estimation Data Annotations

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

Sentence level scores:

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



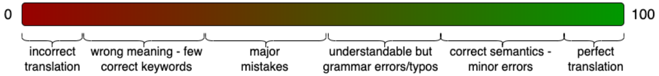
- ▶ Human-targeted Translation Edit Rate (HTER)
  - Minimum number of edits needed to reach from the MT to the post-edited sentence
  - Normalised by sentence length $\longrightarrow$ 0-1 scale

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 $\longrightarrow$ 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 $\longrightarrow$ 0-1 scale

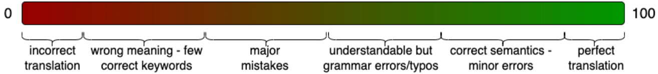
Word level scores:

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 $\longrightarrow$ 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 $\longrightarrow$ 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 $\longrightarrow$ Monolingual: TERcom
  - Source-MT
  - Source-PE

  $\left.\begin{array}{l}\text{Source-MT}\\ \text{Source-PE}\end{array}\right\}$ 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 $\longrightarrow$ Monolingual: TERcom
  - Source-MT
  - Source-PE

  $\left.\begin{array}{l}\text{Source-MT}\\ \text{Source-PE}\end{array}\right\}$ 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 $\longrightarrow$ Monolingual: TERcom
  - Source-MT
  - Source-PE

  $\left.\begin{array}{l}\text{Source-MT}\\ \text{Source-PE}\end{array}\right\}$ 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 $\longrightarrow$ Monolingual: TERcom
  - Source-MT
  - Source-PE

  $\Big\}$ 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 $\longrightarrow$ Monolingual: TERcom
  - Source-MT
  - Source-PE

$\left.\begin{array}{l}\text{Source-MT}\\ \text{Source-PE}\end{array}\right\}$ 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 $\longrightarrow$ Monolingual: TERcom
  - Source-MT
  - Source-PE

  $\left.\right\}$ Bilingual: SimAlign

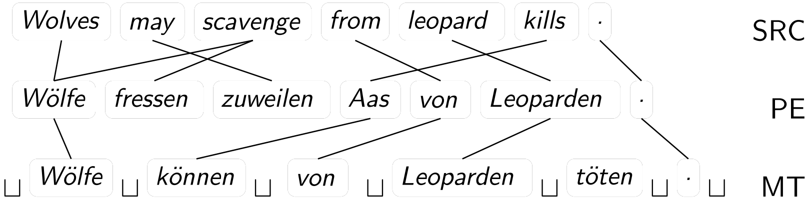
	Wolves		*may*		*scavenge*		*from*		*leopard*		*kills*		*.*	SRC
	Wölfe		*fressen*		*zuweilen*		*Aas*		*von*		*Leoparden*		*.*	PE
␣	*Wölfe*	␣	*können*	␣	*von*	␣	*Leoparden*	␣	*töten*	␣	*.*	␣		MT

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 $\longrightarrow$ Monolingual: TERcom
  - Source-MT
  - Source-PE

  $\left.\begin{array}{l}\text{Source-MT}\\ \text{Source-PE}\end{array}\right\}$ Bilingual: SimAlign

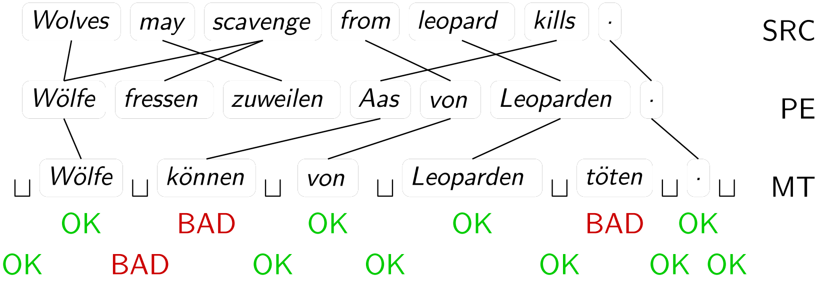


Word-level tag extraction

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

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

  Bilingual: SimAlign

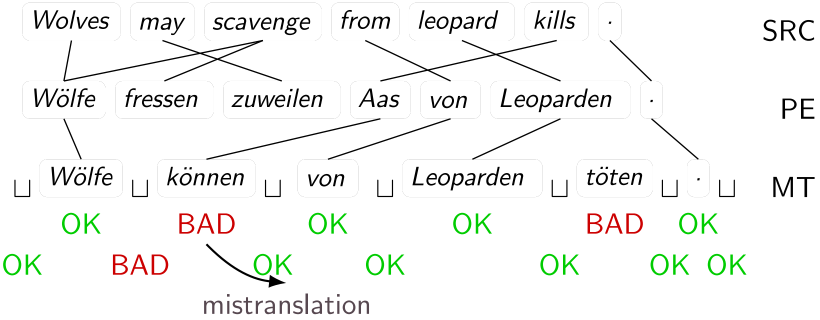


Word-level tag extraction

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

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

$\left.\begin{array}{l}\text{Source-MT}\\ \text{Source-PE}\end{array}\right\}$ Bilingual: SimAlign

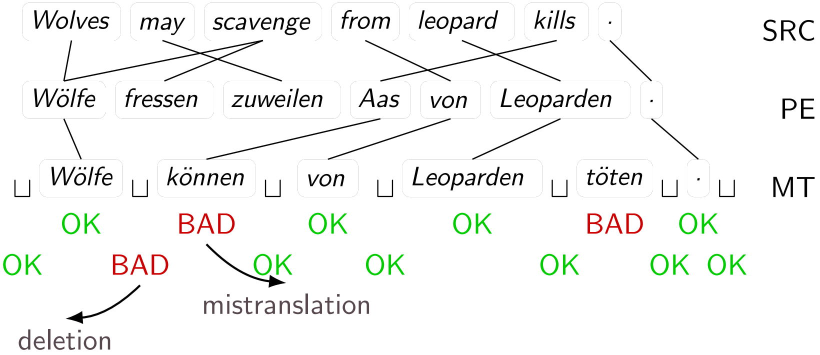


Word-level tag extraction

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

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

  $\Big\}$ Bilingual: SimAlign

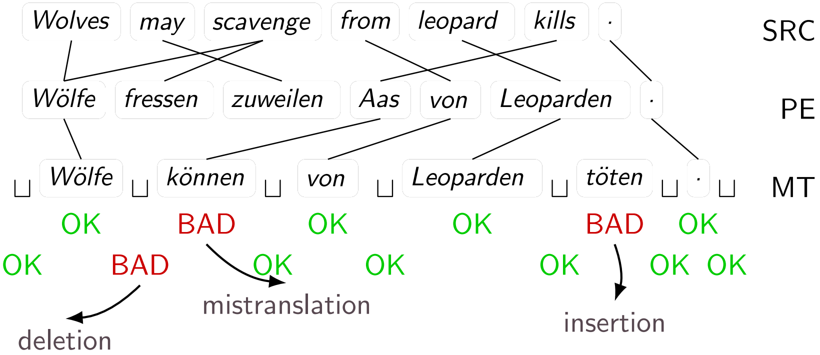


Word-level tag extraction

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

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

  } Bilingual: SimAlign

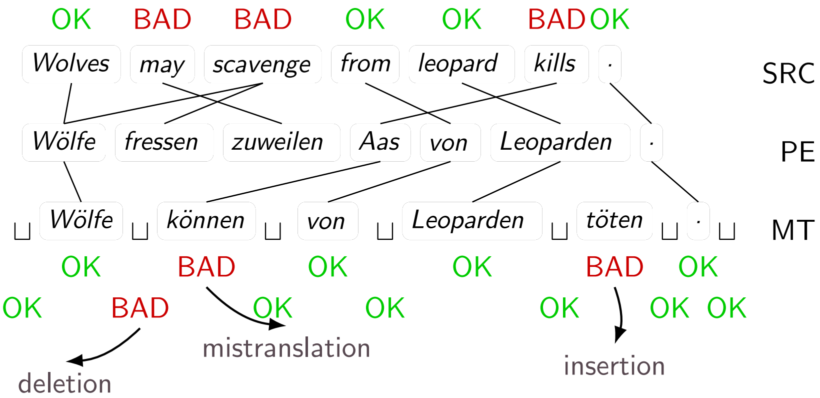


Word-level tag extraction

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

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

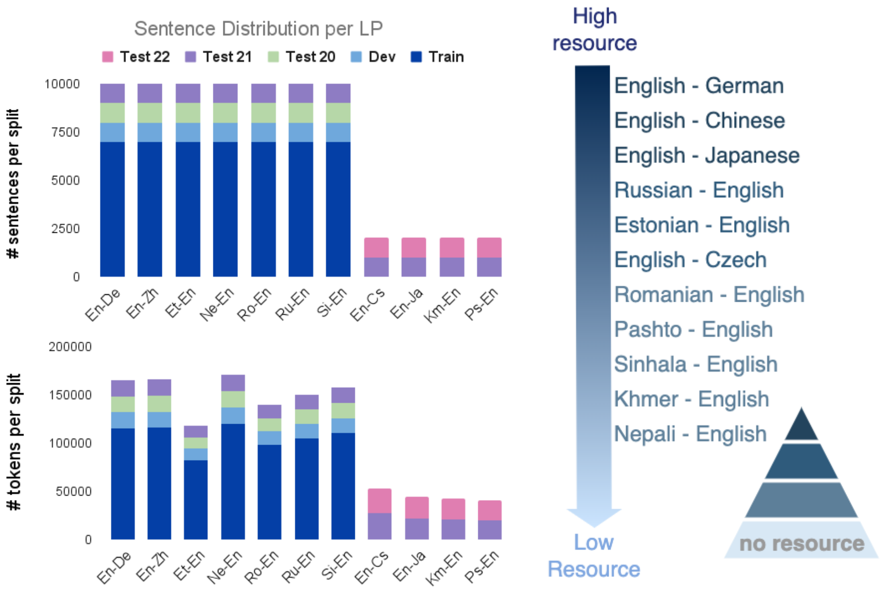
  } Bilingual: SimAlign



Annotations by language pair

Annotations by language pair

Annotations by language pair



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 → STD scores
  - Aleatoric uncertainty
  - Proxy to noisy/complex segments

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 → STD scores
  - Aleatoric uncertainty
  - Proxy to noisy/complex segments

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 → STD scores
  - Aleatoric uncertainty
  - Proxy to noisy/complex segments

motivate **uncertainty aware** approaches

- 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 → STD scores
  - Aleatoric uncertainty
  - Proxy to noisy/complex segments

motivate **uncertainty aware** approaches

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

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 → STD scores
  - Aleatoric uncertainty
  - Proxy to noisy/complex segments

motivate **uncertainty aware** approaches

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

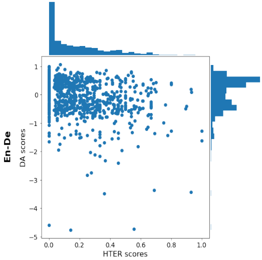
motivate **context aware** approaches

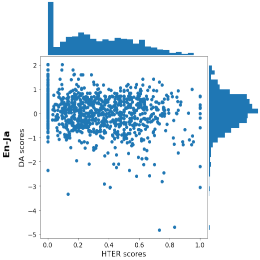
✓ Useful features for quality estimation

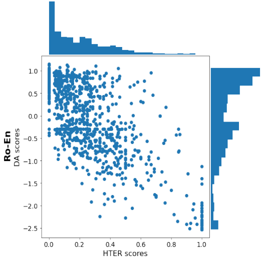
DA-HTER score correlations

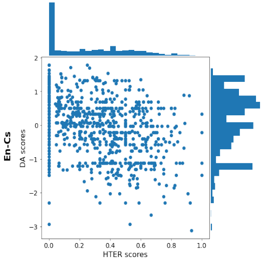
High-resource language pairs:

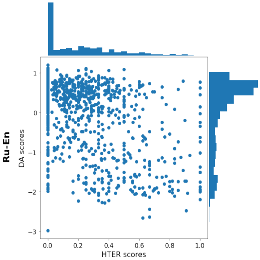
- Different score distributions
- HTER scores (horiz.) are skewed to zero
- Upper-left corner → high-quality translations





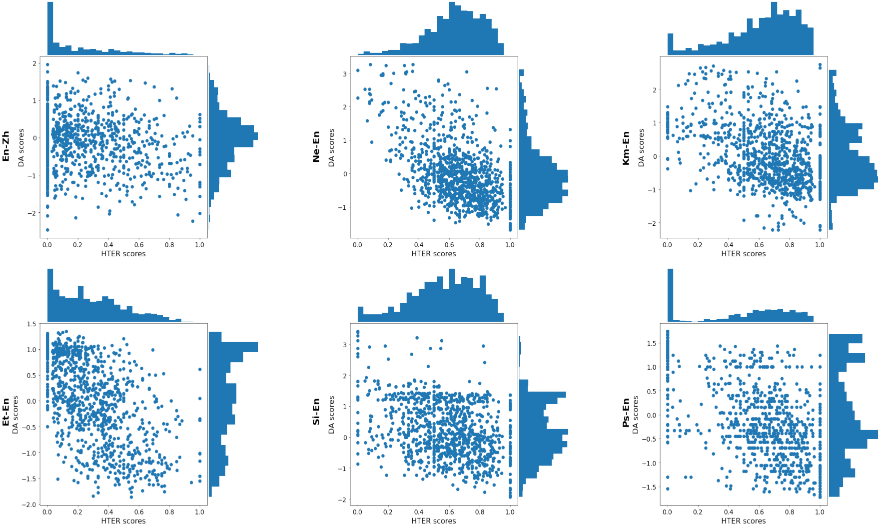






DA-HTER score correlations

Mid & Low resource language pairs:



More on scores and correlations

	Avg. DA ↑	Avg. HTER ↓
En-De	82.61	0.18
Ro-En	69.18	0.24
En-Ja	67.96	0.36
En-Cs	66.94	0.26
En-Zh	62.86	0.23
Et-En	60.09	0.29
Ps-En	53.53	0.53
Si-En	51.42	0.59
Km-En	46.58	0.65
Ne-En	36.51	0.66
Ru-En	68.67	0.23

	Pearson	Spearman
En-De	-0.42	-0.48
Ro-En	-0.76	-0.71
En-Ja	-0.14	-0.11
En-Cs	-0.41	-0.46
En-Zh	-0.21	-0.16
Et-En	-0.61	-0.63
Ps-En	-0.71	-0.67
Si-En	-0.29	-0.28
Km-En	-0.49	-0.43
Ne-En	-0.54	-0.49
Ru-En	-0.51	-0.47

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 $\longrightarrow$ low quality
- HTER score from PE: 0.33 $\longrightarrow$ high quality

Discrepancies

Case 2: Heavy post-editing

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.

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

PE

两人对战陷入僵局，最后双双昏倒。

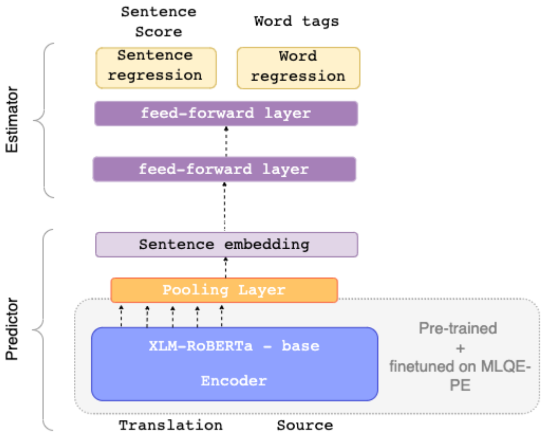
The two people battled to a standstill and both fell into a coma.

- 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

Languages	Pearson r ↑	RMSE ↓
Direct Assessment		
En-De	0.403	0.433
En-Zh	0.525	0.534
Ru-En	0.677	0.492
Ro-En	0.818	0.408
Et-En	0.660	0.543
Ne-En	0.738	0.524
Si-En	0.513	0.626
En-Cs	0.352	0.686
En-Ja	0.230	0.617
Km-En	0.562	0.614
Ps-En	0.476	0.711
AVG	0.541	0.562

Languages	Pearson r ↑	RMSE ↓
HTER		
En-De	0.529	0.129
En-Zh	0.282	0.246
Ru-En	0.448	0.188
Ro-En	0.862	0.111
Et-En	0.714	0.149
Ne-En	0.626	0.160
Si-En	0.607	0.159
En-Cs	0.306	0.206
En-Ja	0.098	0.232
Km-En	0.576	0.196
Ps-En	0.503	0.290
AVG	0.502	0.188

Baseline word-level results

	Words in MT		Words in SRC	
Languages	MCC ↑	F_1-Multi ↑	MCC ↑	F_1-Multi ↑
En-De	0.370	0.415	0.322	0.363
En-Zh	0.247	0.308	0.241	0.295
Ru-En	0.256	0.319	0.251	0.292
Ro-En	0.536	0.553	0.511	0.539
Et-En	0.461	0.512	0.405	0.459
Ne-En	0.440	0.483	0.390	0.438
Si-En	0.425	0.456	0.335	0.379
En-Cs	0.273	0.372	0.224	0.312
En-Ja	0.131	0.217	0.175	0.272
Km-En	0.351	0.409	0.279	0.355
Ps-En	0.313	0.425	0.249	0.361
AVG	0.346	0.402	0.307	0.370

In practice

Already used:

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

That's not all

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 :)

Thank you!

THANK YOU!

Questions?

arxiv paper

github code