A language comparison of Human Evaluation and Quality Estimation

Silvio Picinini - eBay
Adam Bittlingmayer - ModelFront

ebay

MT Quality

Human Evaluation

Quality scores by human linguists

Quality Estimation

Quality scores by machine

No reference translation used

Also called “confidence score” and “risk prediction”

Aggregated for automatic quality *evaluation*

ebay

Goals

“How does machine QE correlate with human evaluation?”

- Compare line-level and aggregate numbers

“What causes differences between QE and human evaluation?”

- Analyse QE line-level issues
  - Get insights for QE

ebay

Human Evaluation

The content:

- 200 segments
  - Various lengths
  - With and without placeholders/tags
- 4 MT outputs per language - one customized
- 2 languages - pt-BR and es-CO

Three expert evaluators per language - reliable results

Scores range from 1 to 4 stars - normalized to 0-100

ebay

Quality Estimation

Generic production system - no custom data, no locales, no context, used for many use cases

Originally a Risk Prediction (0% good, 100% bad), which includes *source-side ambiguity*

Risk is reversed to become QE score (0 bad, 100 good)

Very convenient, but challenging for the QE system to match humans operating with many more inputs.

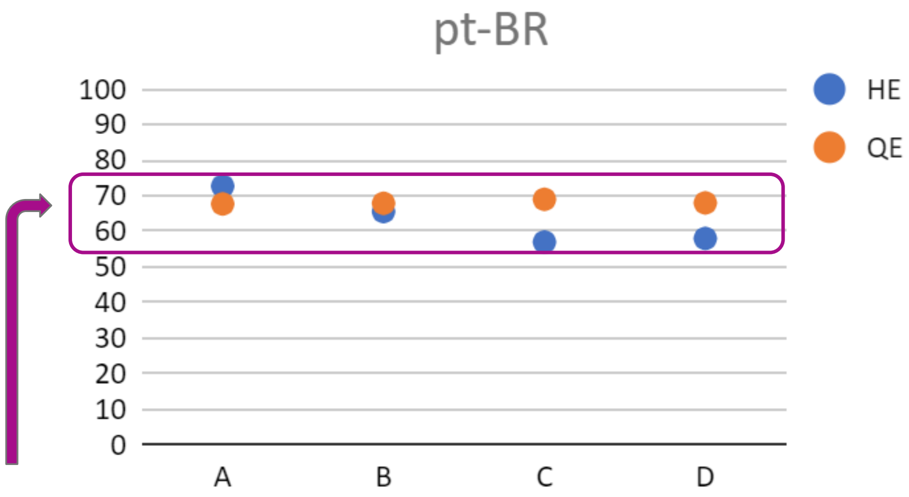
ebay

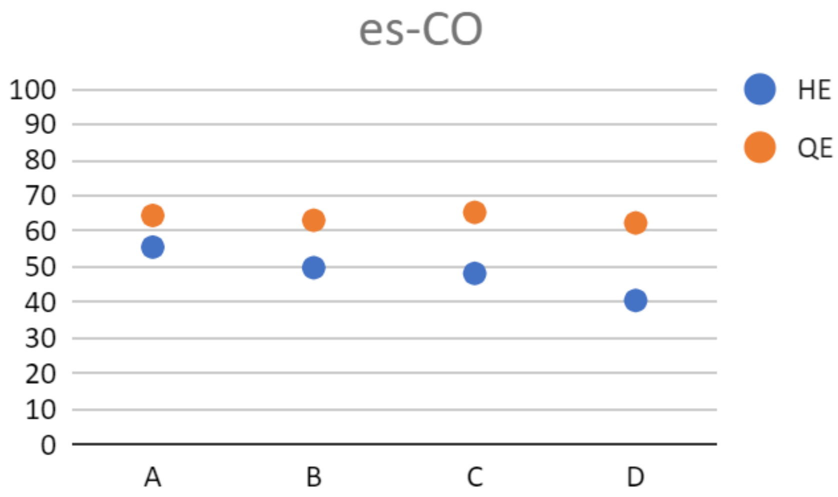
Numbers

Comparison for the set

QE for pt-BR was closer to the HE, es-CO was a little further

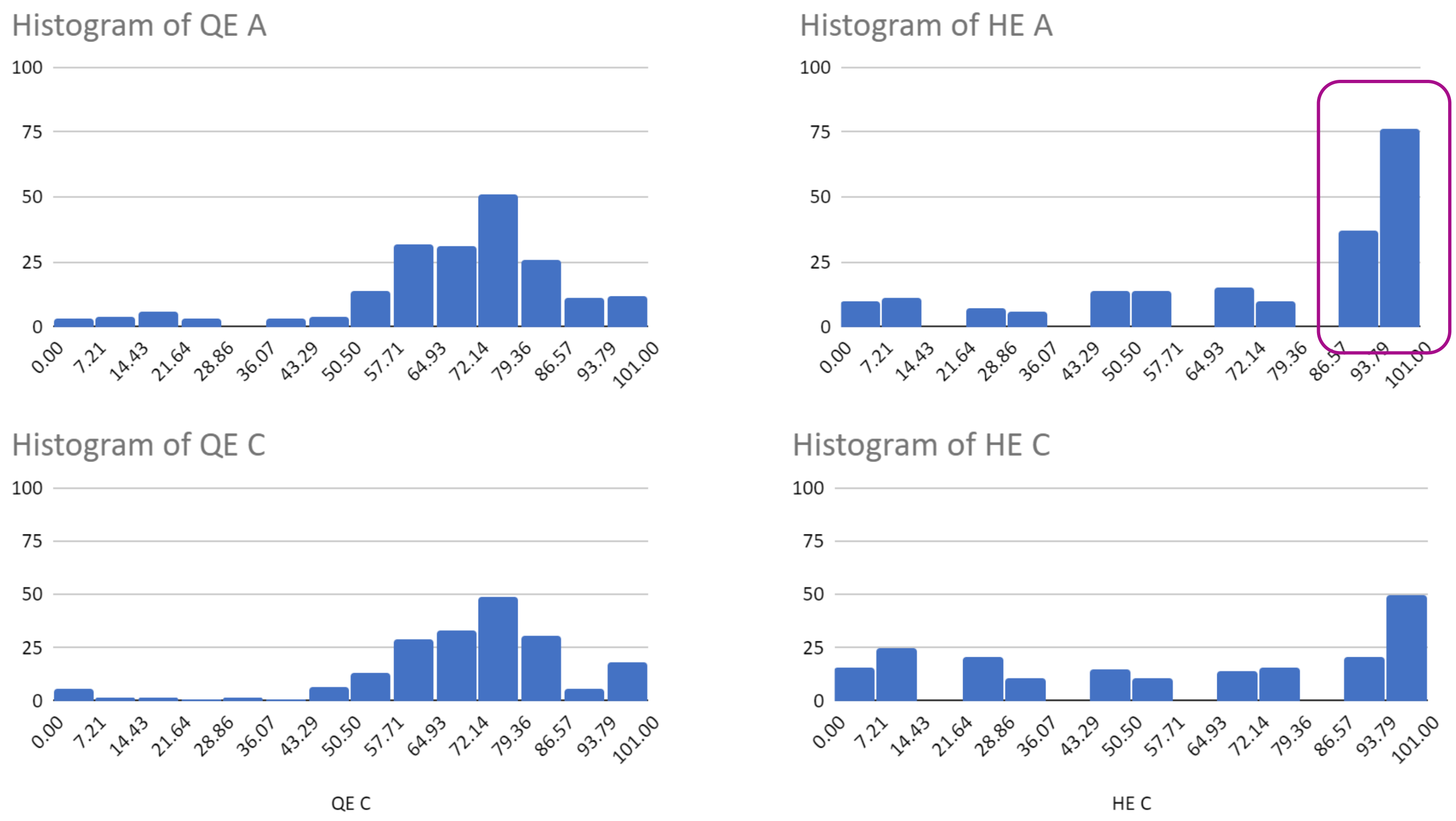
QE was close to the best HE and further away for the worse





The QE is within a narrow range close to the HE. This is a good result.

Comparison QE HE histograms - pt-BR


Histogram of QE A

Histogram of HE A

Histogram of QE C

Histogram of HE C

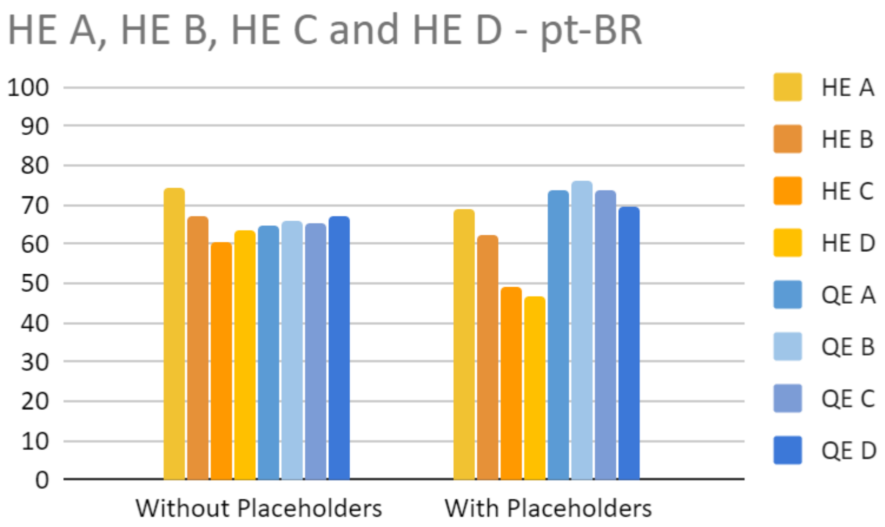
QE better for best MT

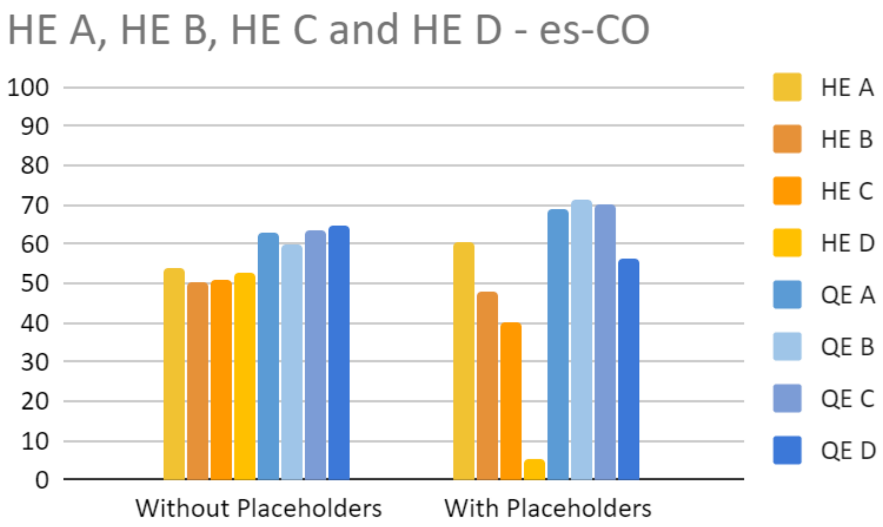
Mostly misses the concentration of near perfect.

Comparison for Placeholders

HE had lower scores for segments with placeholders - in { } format

QE had higher results with placeholders

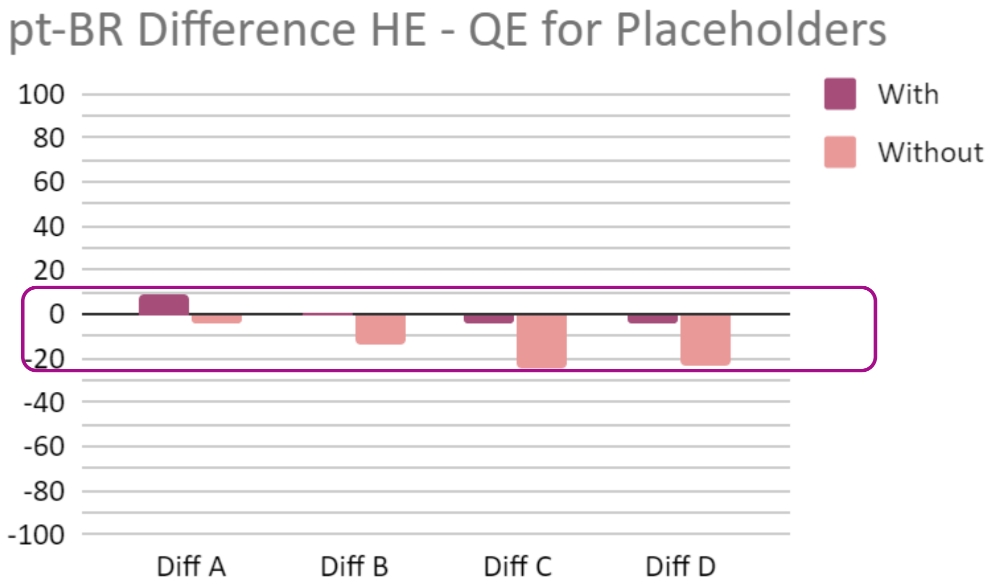




Comparison for differences HE - QE

QE in general overestimates the quality (most differences are negative)

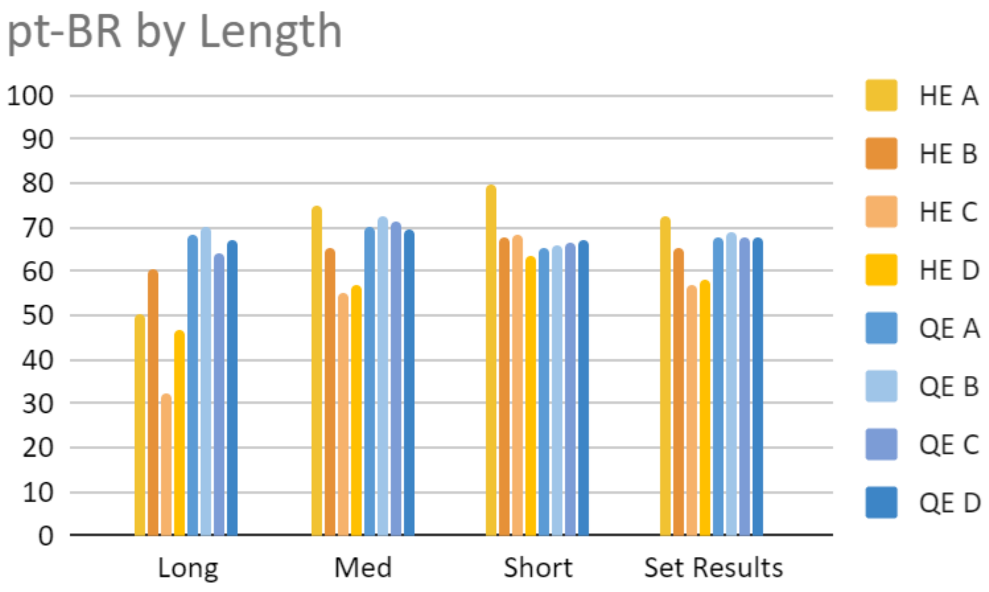
The worst the HE, the greater the difference to the QE

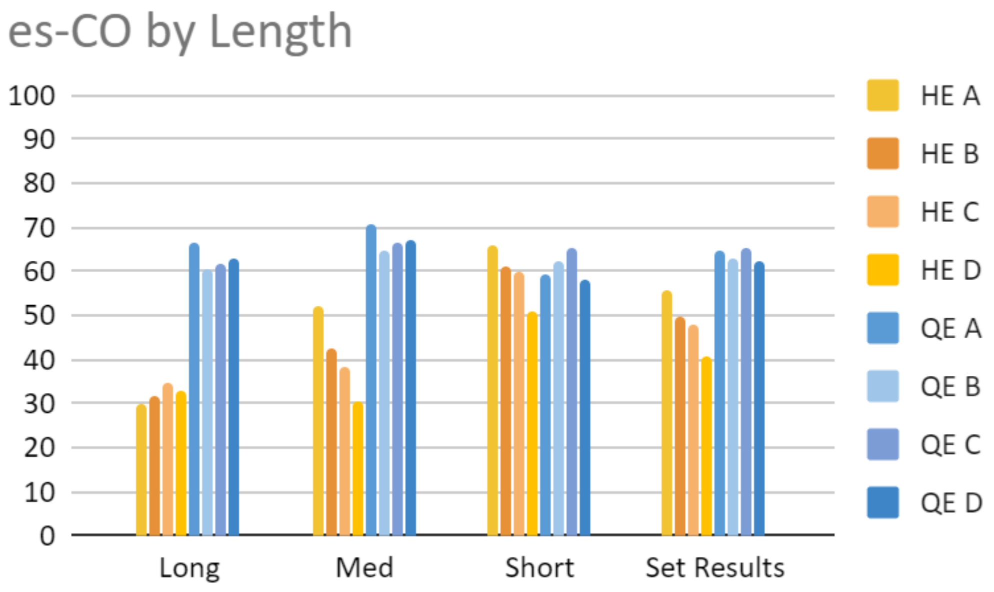


Comparison for Length

HE clearly scored Long < Med < Short

QE did not differentiate, but results for Short are close





Language Issues

ebay

Language Issues

Where we looked:

- HE is much higher than QE - QE **underestimates** quality

- HE is much lower than QE - QE **overestimates** quality

- HE has a wide range of values among the 4 MT outputs (shows **varied translations**, from good to bad)

ebay

Examples

If we don't hear back, this request will be closed on {1} and the hold on this transaction will be removed.	Si no recibimos una respuesta, esta solicitud se cerrará el {1} y se borrará la cuenta de esta transacción.

Mistranslation: the meaning changed from “the hold on the transaction will be removed” (positive) to “the account will be erased” (negative).

HE:11
QE:89

The HE noticed that but the QE did not.

If we don't hear back, this request will be closed on {1} and the hold on this transaction will be removed.	Si no recibimos respuesta, esta solicitud se cerrará el {1} y se eliminará la retención de esta transacción.

Examples

Click to learn about Top Rated Sellers	Clique para saber mais sobre vendedores nível Top

Glossary SRC	Glossary TGT
Top Rated Seller	Vendedor nível Top

Terminology: eBay has a specific terminology for “Top Rated Seller”, which includes the use of an “untranslated” word Top.

HE:92
QE:8

The QE may see this as a possible defect and rate the translation low. HE is aware that in our context the translation is perfect.

Examples

Silver shooting star for feedback score from 1,000,000 or more	Estrela de tiro de prata para a pontuação de feedback de 1 milhão ou mais

HE:11
QE:66

Idioms and figurative meaning:
The expression "shooting star" was translated as "a star of the activity of shooting a gun".

Basketball? Also a shooting star.

ebay

Examples

{1}Not a registered user{2}	{1} Não é um utilizador registado {2}

Locale: One MT is more influenced by data from European Portuguese. The MT above contains two examples of that.

HE:0
QE:81

{1}Not a registered user{2}	{1}Não é um usuário cadastrado{2}

Examples

A decision has been made about the dispute that was filed by {1}.	Uma decisão foi feita sobre A disputa que foi registrada em {1}.

Placeholders: they introduced an ambiguity for the MT, which was clear for HE. The source says “filed by {1}” and it means “filed by a person”. The MT and the QE thought that it meant “filed by this date”.

HE:0
QE:97

A decision has been made about the dispute that was filed by {1}.	Foi tomada uma decisão sobre a disputa que foi apresentada por {1}.

Examples

PostePay	Envío postal
PostePay	PostePay

Untranslatable:

The name of a service was translated as “Postal shipping”.

The HE noticed that, the QE somewhat.

HE:0
QE:24

Examples

We're aware of this issue and are working to fix it as soon as possible.	Este problema y estamos tratando de solucionar el problema lo antes posible.

Omission:

The translation just says “This issue”, omitting “We’re aware of”.

HE:11
QE:96

We're aware of this issue and are working to fix it as soon as possible.	Somos conscientes de este problema y estamos trabajando para solucionarlo lo antes posible.

ebay

Language Issues

What are some of the reasons for discrepancy between HE and QE?

- Mistranslations not recognized
- Terminology
- Idioms and figurative meaning
- Locale
- Placeholders
- Untranslatables
- Omissions

ebay

Takeaways

- The main generic QE system has aggregate scores in a similar range as HE. This is promising.
- Customization is key to QE for evaluation, to shape the output to the custom translation and evaluation guidelines
- Findings in custom data can help improve accuracy on non-custom errors
- QE is a rising technology that will be widely present in many MT uses in the near future

Future step: Use a trained engine

Acknowledgements

Our thanks to the language experts that worked on this:

Melany Laterman and Patricia Lawler

eBay Language Specialists

for Brazilian Portuguese and Latin American Spanish

ebay

Questions?

Thanks
Obrigado
Gracias

ebay