PEMT human evaluation at 100x scale with risk-driven sampling

by Kirill Soloviev

CEO & co-founder

about me
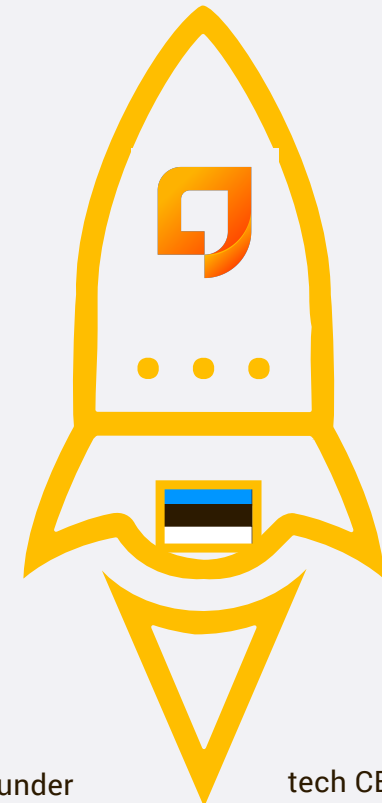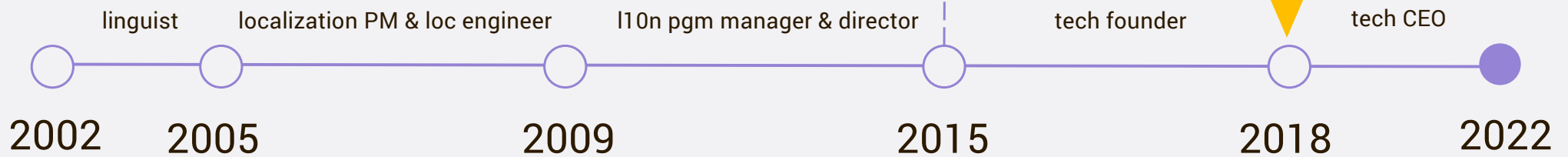
37 years old

20 years GILT

7 years TQM
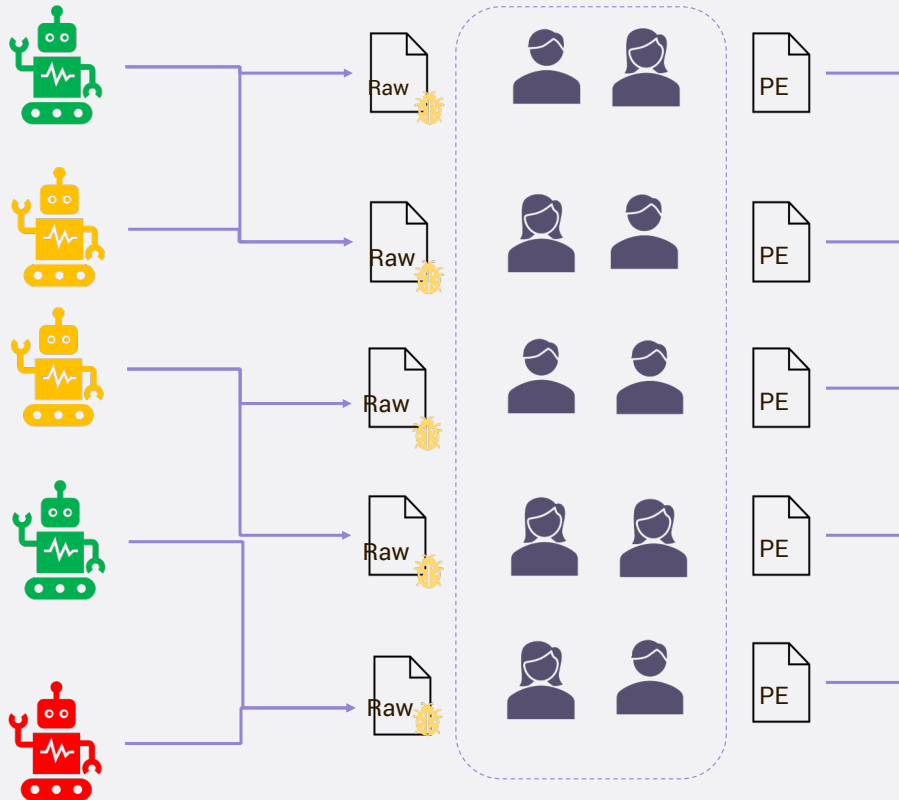
linguist   localization PM & loc engineer   l10n pgm manager & director   tech founder   tech CEO

2002   2005   2009   2015   2018   2022

human evaluation for actionable insights

page
04

Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas, Orlando, USA, September 12-16, 2022. Volume 2: Users and Providers Track and Government Track   *Page 4*
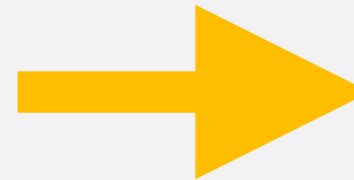
non-scientific matrix of MT quality evaluation approaches

Not shown to scale!!!

High cost

human edit distance (subsegment) ***

human MQM (subsegment)

Low detail

human adequacy - fluency (segment)

High detail

automatic w/reference (BLEU)

automatic non-reference (COMET)

Low cost

page 05

Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas, Orlando, USA, September 12-16, 2022. Volume 2: Users and Providers Track and Government Track

Page 5

output of human quality evaluations

Rating Scale   Error Annotation

English (en-US)	Estonian (et-EE)	Adequacy	Fluency	MQM Annotation	Edit Distance
The quick brown fox jumps over the lazy dog	Kiire pruun rebane hüppab üle laisa kaerakoera	4 out of 4	4 out of 4	Accuracy / Major Spelling / Minor	5%
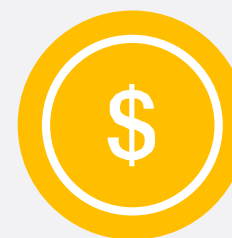
risk-driven sampling
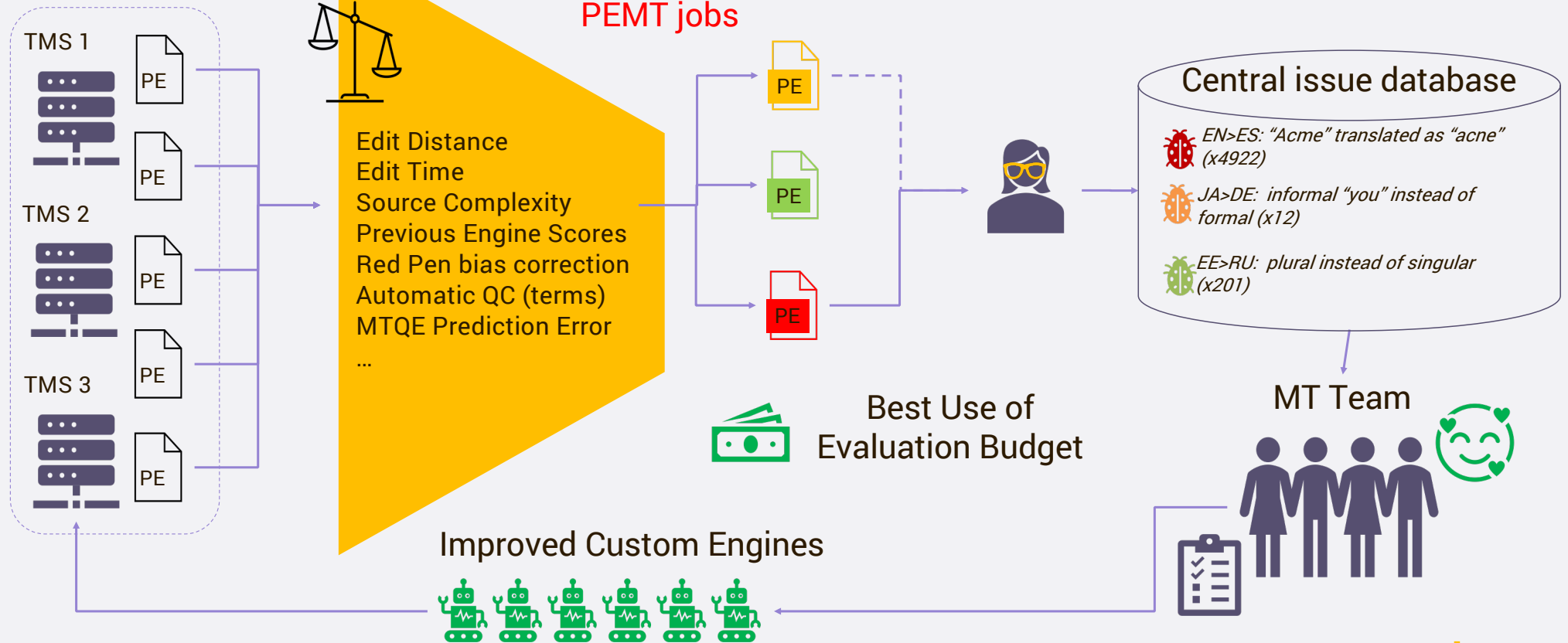
Continuously calculated ED / MTQE

Auto risk scoring and budget-limited sampling of PEMT jobs

Human evaluation

Specific, actionable MT quality error reports

TMS 1

PE
PE

TMS 2

PE
PE

TMS 3

PE

Edit Distance
Edit Time
Source Complexity
Previous Engine Scores
Red Pen bias correction
Automatic QC (terms)
MTQE Prediction Error
...

PE
PE
PE

Best Use of Evaluation Budget

Central issue database

🐞 *EN>ES: "Acme" translated as "acne" (x4922)*

🐞 *JA>DE: informal "you" instead of formal (x12)*

🐞 *EE>RU: plural instead of singular (x201)*

MT Team

Improved Custom Engines
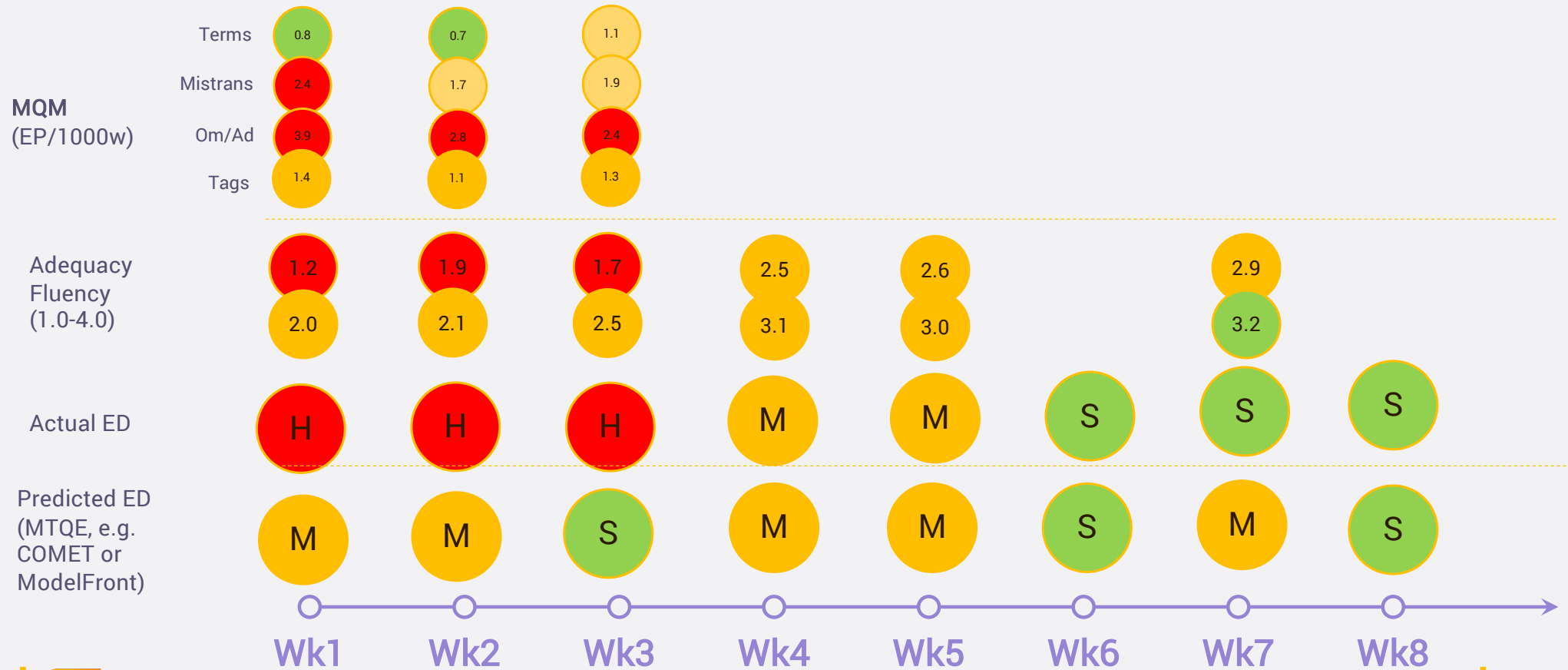
example quality
risk rules

PEMT jobs should be more likely to be picked for human quality evaluation when:

- [x4] Average Edit Distance for language X changed >10% over last 1 month
- [x2] Maximum Edit Distance for engine Y hit 60% twice over 10 translators
- [x0.5] Predicted Edit Distance for MTQE model Z differed >30% from ED
- [x1.5] Post-editor M's median Edit Distance is >30% different from average
- etc.

+ Budget-guided cutoff point

continuous q.evaluation blueprint for PEMT

	Wk1	Wk2	Wk3	Wk4	Wk5	Wk6	Wk7	Wk8
MQM (EP/1000w) — Terms	0.8	0.7	1.1					
Mistrans	2.4	1.7	1.9					
Om/Ad	3.9	2.8	2.4					
Tags	1.4	1.1	1.3					
Adequacy / Fluency (1.0-4.0) — Adequacy	1.2	1.9	1.7	2.5	2.6		2.9	
Fluency	2.0	2.1	2.5	3.1	3.0		3.2	
Actual ED	H	H	H	M	M	S	S	S
Predicted ED (MTQE, e.g. COMET or ModelFront)	M	M	S	M	M	S	M	S

benefits

Improved ROI on human evaluations

Faster custom engine improvement

More reliable fix of Red Pen Syndrome

thank you!
time for Q&A

Email: Kirill.Soloviev@contentquo.com
LinkedIn: https://www.linkedin.com/in/kirillsoloviev/