RWS Moravia

RWS Moravia Operationalizing Machine Translation Quality Estimation (QE)

Miklós Urbán,
Senior Solutions Architect

Maribel Rodríguez,
Language Technology Deployment Manager

www.rws.com/moravia

Agenda

- Introduction
- Methodology
- Potential Business Cases
- Technical Setup
- Conclusions

RWS Moravia

Quality estimation is a method used to automatically provide a quality indication for machine translation output without depending on human reference translations. In more simple terms, it's a way to find out how good or bad the translations are that are produced by an MT system without human intervention.

Yet in the machine translation space, there's evidence to show that good quality estimation eases the burden on human editors. With an automated system that highlights mistakes before the human process even begins, the editors can zero in on the areas of a piece of content that most likely need attention.

https://www.forbes.com/sites/forbestechcouncil/2019/01/24/why-quality-estimation-is-the-missing-link-for-machine-translation-adoption
https://tech.ebayinc.com/engineering/machine-translation-the-basics-of-quality-estimation/

The Challenge

With so many different approaches to QE out there and so many variables, **how can we**:

- Evaluate QE performance for different QE options, customers, content types, languages, etc.?
- Identify the business cases that could bring value to RWS Moravia and our clients?
- Figure out when is the right time to implement QE in a specific workflow?
- Continue to monitor the performance of QE after it has been implemented?

RWS Moravia

How to Evaluate QE?

RWS Moravia

Common Methodology

Pre-translate the content using MT

Obtain both pre-production MT QE and post-production TER scores

Compare QE score with actual TER score

Analyze the results

Considerations:

- Post-editors are not exposed to QE
- QE initially runs in the background
- Production may apply different workflows
- Translation is not analyzed for over-editing or under-editing

RWS Moravia

Input Metric | Quality Estimation

- Quality estimation is available from multiple sources
- QE is based on machine learning algorithms
- To make results comparable, we convert QE results to a 4-choice numeric score system

  - **100%** means QE predicts good raw MT quality
  - **67%** means QE predicts some editing is needed
  - **33%** means QE predicts more editing is needed
  - **0%** means QE predicts poor raw MT quality

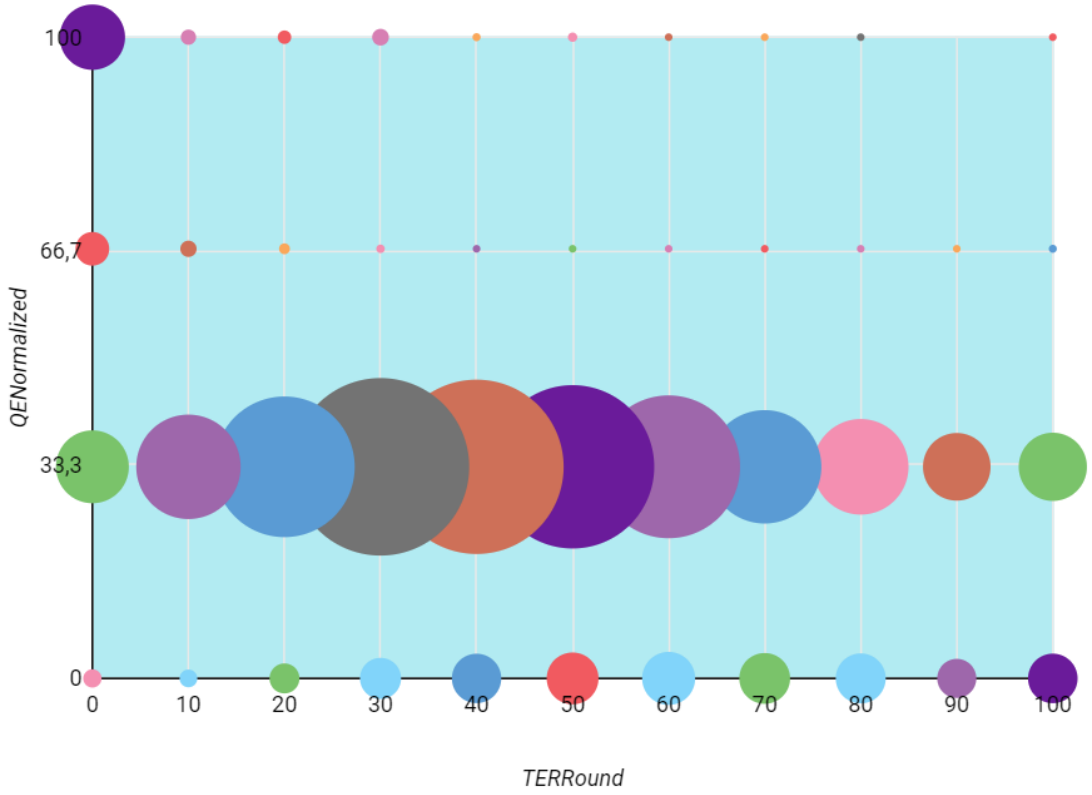
RWS Moravia

Input Metric | Translation Edit Rate (TER)

- Suited to quantify the post-editing effort
- RWS Moravia has been using TER in production for over a decade
  - Number of edits needed to modify raw MT to produce a final translation
  - TER = edits / reference word count
    - where edits = insertions, deletions, substitutions and shifts
  - The closer the score is to 0, the less post-editing effort is assumed
- We round TER scores to multiples of 10%

RWS Moravia

How to Use the Data?



Each segment can be plotted on a chart

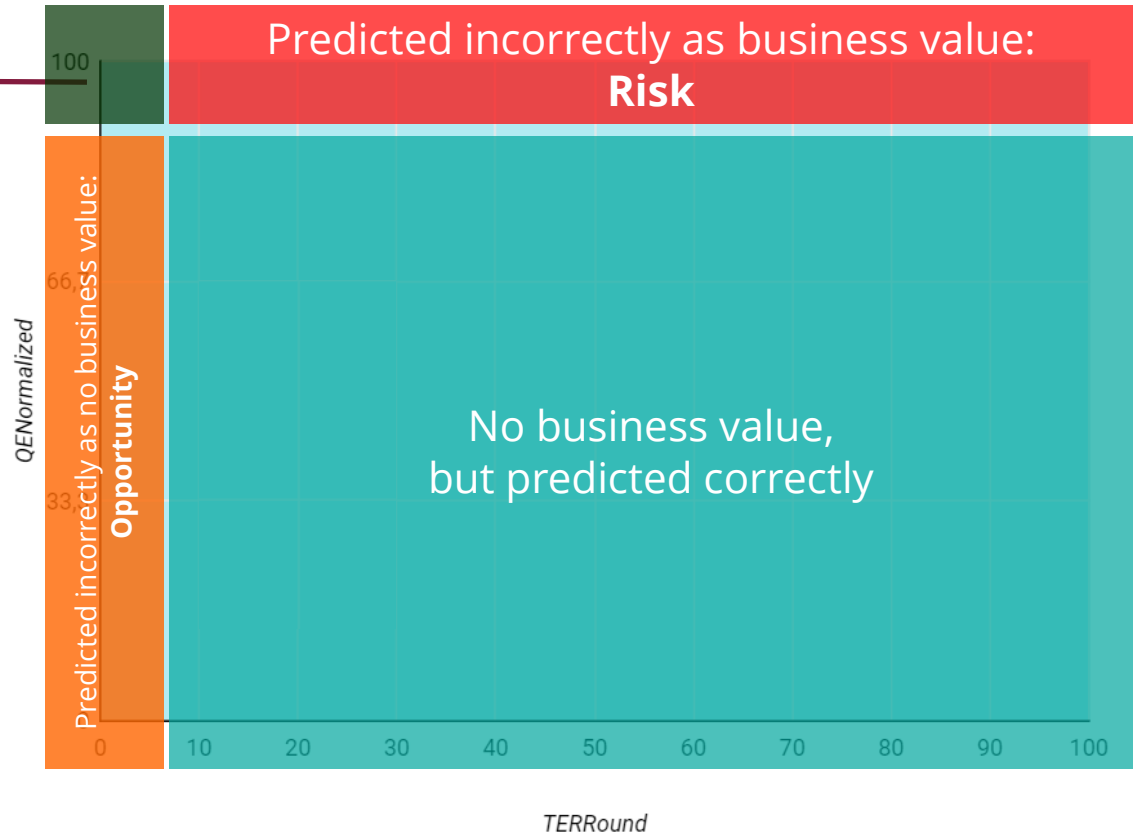
We created a bubble chart with:

- Y-axis: QE score
- X-axis: TER score
- Size of bubble: number of segments

RWS Moravia

Interpreting the Bubble Chart

Predicted correctly, adding business value



Based on the business case, we can identify areas of the chart driving the business value.

RWS Moravia

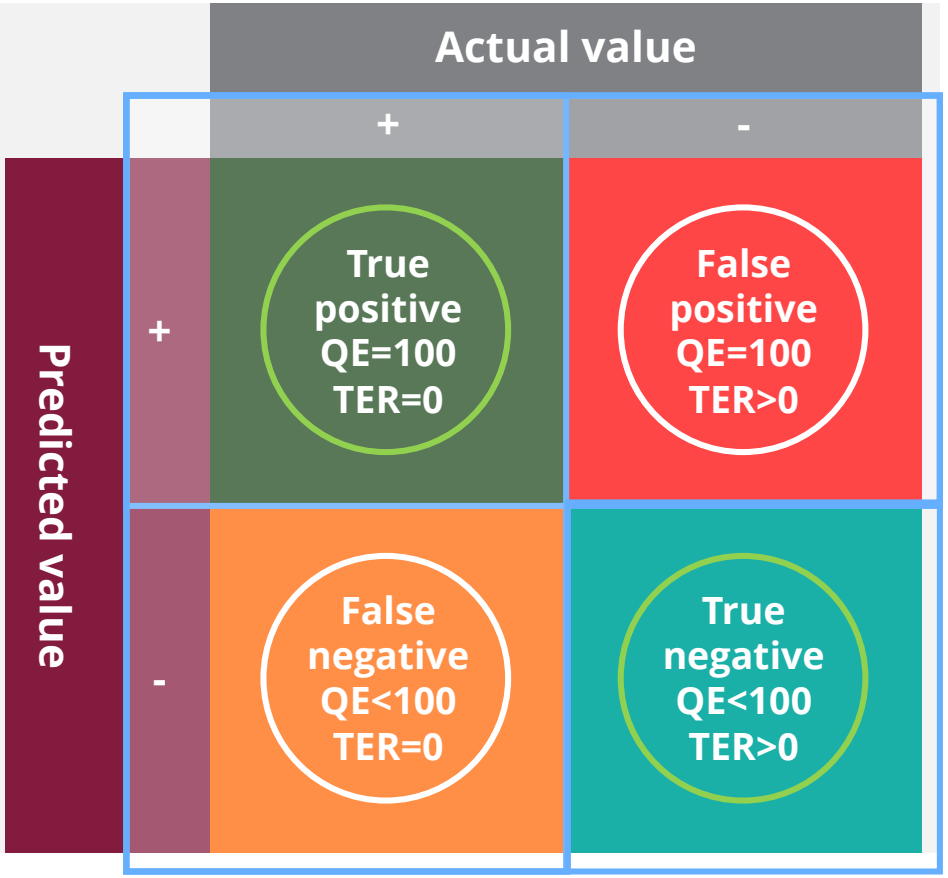
Output Metrics

$$\text{Accuracy} = \frac{\text{True Positive} + \textbf{True Negative}}{\text{Total}}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \textbf{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \textbf{False Negative}}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



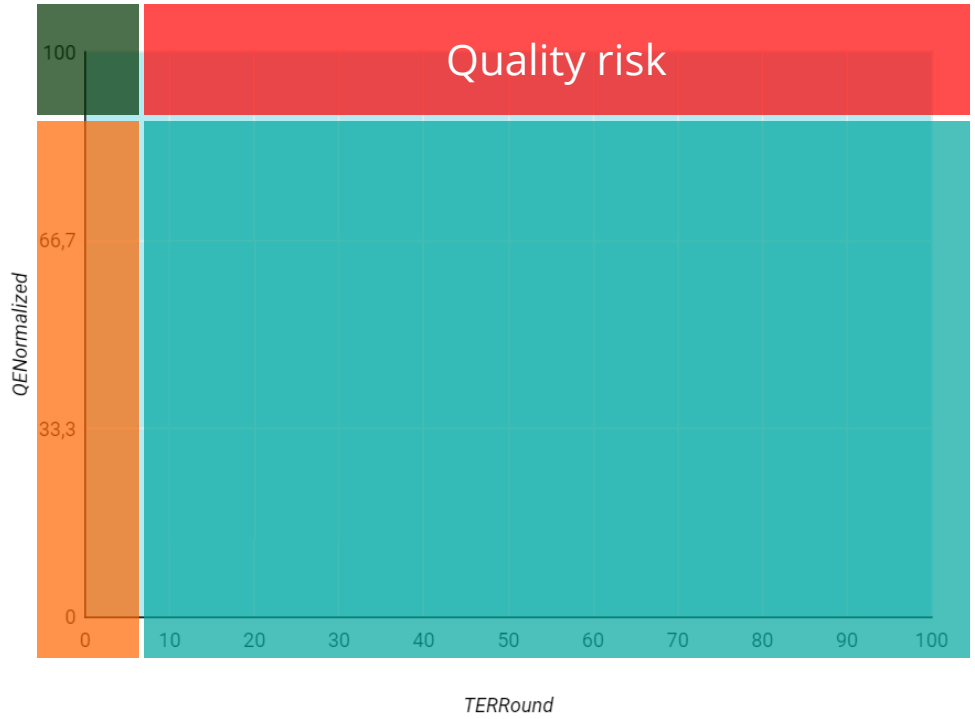
RWS Moravia

What Are the Business Cases?

RWS Moravia

Take Advantage of Good MT Segments

- Eliminate post-editing or apply a light post-editing workflow for good raw MT segments (up to 30% of segments)
- Quality risk for false positives
- We expect a high proportion of non-edited segments to be identified, keeping the quality risk close to zero



RWS Moravia

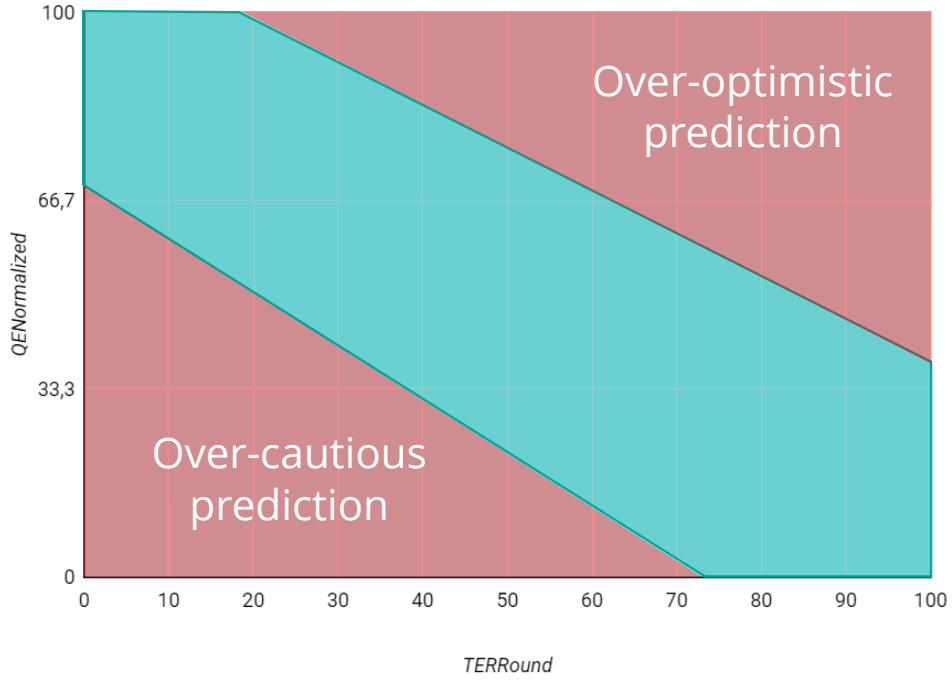
Remove Burden of Reading Poor MT

- Does it really increase productivity?
- Risk of deleting good MT
- We expect a high proportion of poor-quality raw MT to be discarded with minimal loss of good MT

RWS Moravia

Assessing MT Quality and Applying Fair, Pre-production Pricing

- A good accuracy could allow MT quality measurement without human reference
- High accuracy (95+%) could allow pricing to be based on QE



RWS Moravia

Road to Operationalization

RWS Moravia

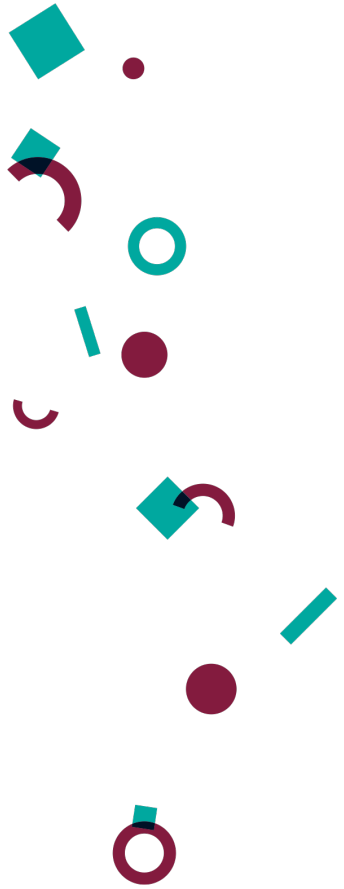
Considerations

- Multiple QE sources
  - Choose the best option that fits our purposes
- QE performance may depend on multiple factors
  - Language pair
  - Client
  - Content type
- We need to establish reproducible metrics that can be measured over a large sample

RWS Moravia

Pilot

- No. of customers: **1**
- Content types: **2**
- Language pairs: **17**
- Experiment duration: **8 months**



RWS Moravia

Pilot Results

- Methodology enabling:
  - Consistent evaluation of QE technology and tracking its progress
  - Monitoring results against preset thresholds before going live
- Automated dataflow solution
  - Evaluation of usability of different QE systems
  - Data insights through dashboards
- Findings
  - Dependency of QE performance across languages and content types
  - Technology still evolves and shows improved performance over time

RWS Moravia

How to Set up Continuous Tracking?

RWS Moravia

Technical Setup

After populating raw MT into the CAT tool, the QE prediction was run and scores were stored

RWS Moravia

Technical Setup

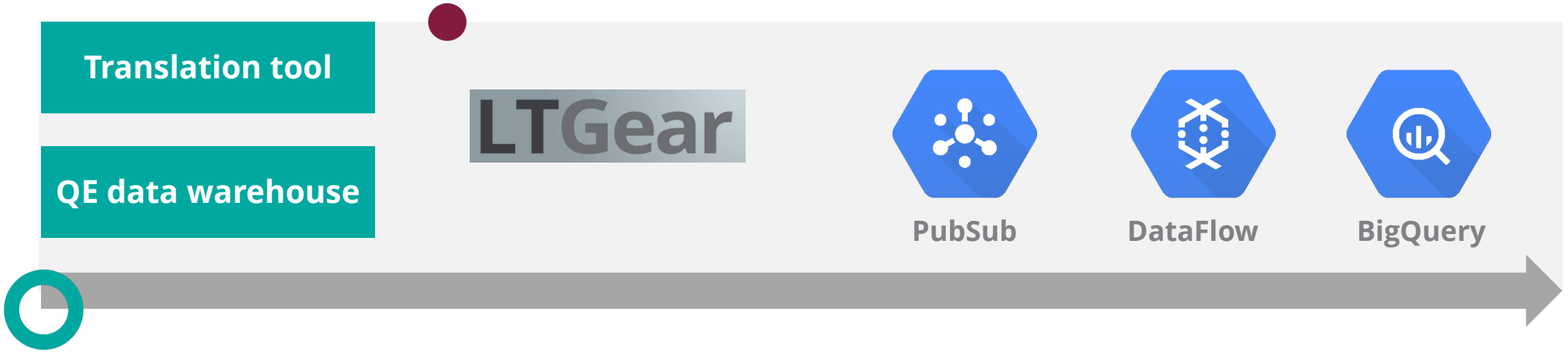
Post-editors completed
the task in the translation tool
without being exposed to QE

RWS Moravia

Technical Setup

We created a streaming data solution that:

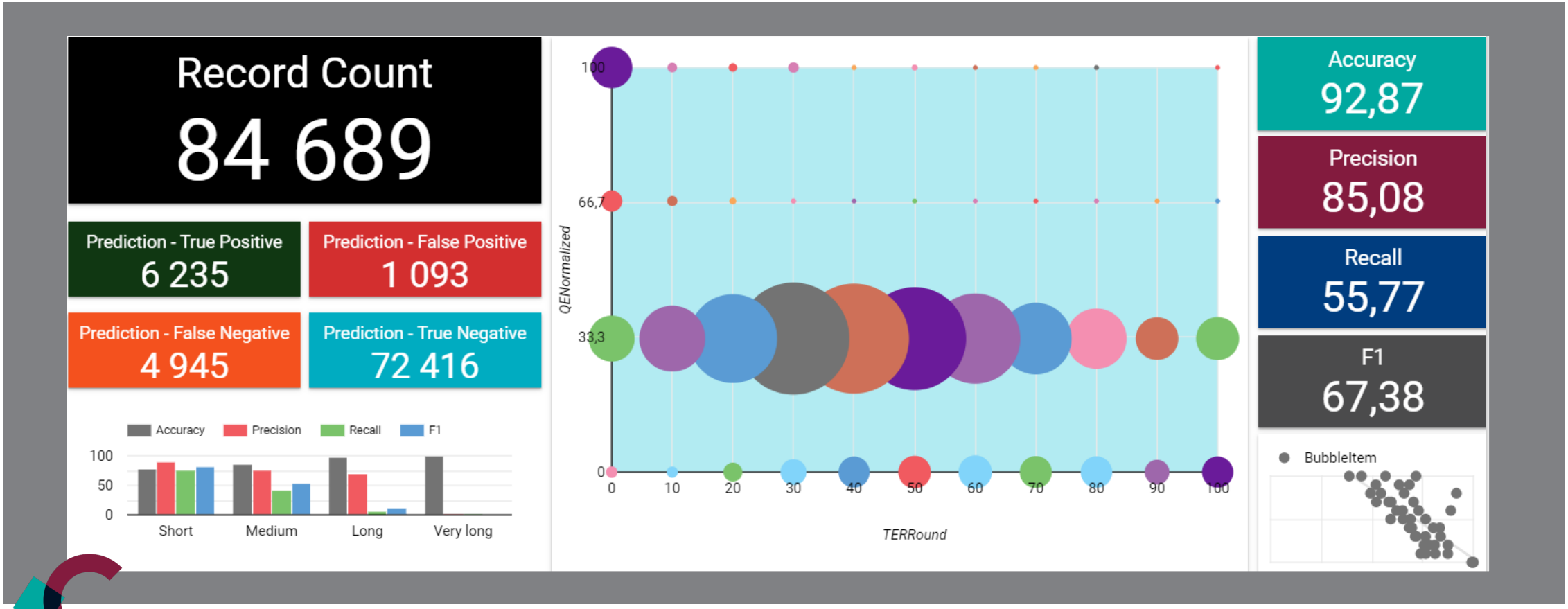
- Takes segment data from the production environment
- Runs the segment through the TER score evaluation in our proprietary software, LTGear
- Matches the QE data stored earlier for the segment
- Streams this data into the bigdata infrastructure of Google Cloud



Technical Setup

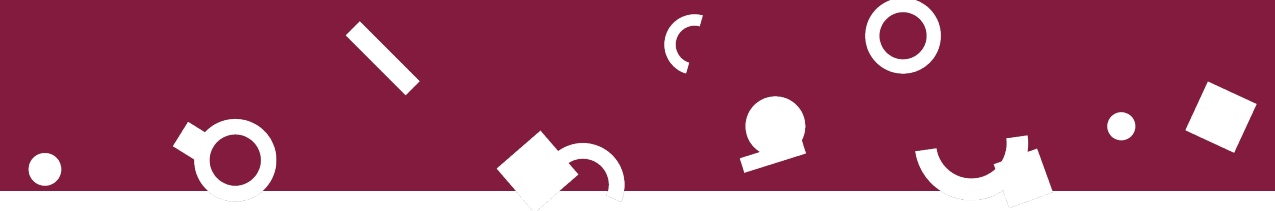
- For each segment, we store its coordinates, timestamp, language pair, client and domain metadata and the MT QE and TER results in BigQuery
- Google Data Studio dashboards help us track and analyze the results

Sample Dashboard



RWS Moravia

Conclusions



RWS Moravia

Conclusions

- QE shows improvement over time and is approaching production readiness in a large LSP setting
- QE performance is highly dependent on language pair and content type
- Robust solution to track performance of QE predictions against post-production metrics is needed
- Thanks to the framework we have put in place, we now have the means to easily monitor the aggregated data in a continuous stream and compare the performance of multiple QE sources

RWS Moravia

Some Questions We Are Really Eager to Answer

- Does it make sense to include segment length beside QE to refine the precision of predictions?
- Is quality retained for high-ranking QE segments that will likely get less attention?
- Do post-editors start from scratch for low-ranking QE segments?
- Is productivity enhanced compared to a workflow without QE?

RWS Moravia

Acknowledgement to All the Team Members that Participated in This Research

Tomáš Burkert
Solutions Architect

Tomáš Fulajtár
MT Researcher

Miklós Urbán
Senior Solutions Architect

Maribel Rodríguez
Language Technology
Deployment Manager

RWS Moravia

Q&A

Thank you