Auto MT Quality Prediction Solution and Best Practice

York Jin & Martin Xiao

Oct. 2020
Agenda

01 Program overview
02 Data collection and model training
03 Perfect MT scenario
04 Inference acceleration
05 Future works
Program overview
Why prediction is needed

Perfect: PE% = 0
Good: 0 < PE% < 20%
Bad: PE% > 20%
## Program overview

<table>
<thead>
<tr>
<th>POC</th>
<th>Model fine-tune</th>
<th>Deploy to Stag.</th>
<th>Deploy to Prod.</th>
<th>Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data collection</td>
<td>Added source English as feature</td>
<td>Model validation by real data</td>
<td>Model size reduction</td>
<td>Perfect MT scenario</td>
</tr>
<tr>
<td>Data washing</td>
<td>(English + MT as input, PE as the label)</td>
<td>Load balancing by using multiple instances</td>
<td>Further validate the model by running pilot projects</td>
<td></td>
</tr>
<tr>
<td>Conceptual design</td>
<td>Added regression to get linear output</td>
<td>Model performance against 25,000 new words project (5-10 in average)</td>
<td>Deploy trained models in DECC (CPU only, with load balancer)</td>
<td></td>
</tr>
<tr>
<td>Result analyze</td>
<td>Generated training data from fuzzy/ICE</td>
<td>Quality index invented and patent applied</td>
<td>OpenVINO inference accelerator (CPU)</td>
<td>RAW MT quality auto scoring scenario</td>
</tr>
<tr>
<td>Prove of concept</td>
<td>Validation framework (correlation scatter diagram with actual PE etc.)</td>
<td></td>
<td>Tensor RT ML inference accelerator validation (GPU)</td>
<td>Engine quality auto evaluation scenario</td>
</tr>
</tbody>
</table>

- Exclusion rules
- Integration with TMS
Data collection and model training

Automated Data collection (Source, MT, human translation, PE)

Data cleaning

Evaluation

Fine-tune

Integrate into production
Data collection and model training

Algorithm – Transformer
Perfect MT scenario

- NMT Quality Prediction Service
  - Source & MT ➔ PS
  - PS ➔ TMS
  - TMS ➔ MT Middleware
  - MT Middleware ➔ MT Engine 1, MT Engine 2, Other MT Engines

- PS: Prediction Score
- PQI: Prediction Quality Index

- PS ≤ PQI ➔ Take as 100% match

- Rule-based Exclusion: False

- 10% - 20% cost saving
- PQI: Patent applied
Perfect MT Scenario

Overall accuracy

![Bar chart showing overall accuracy of recent 2 months pilots for different languages: Chinese (98.30%), French (93.20%), Japanese (98.10%), Korean (95.60%), Spanish (92.30%), German (87.20%).]
Perfect MT Scenario

![Graphs showing MT performance for different projects and languages]
Perfect MT Scenario

Typical prediction failure example

DE golden MT that human linguist marked as “bad“:

Source:

Directory sync is handled by the connector component of the service and can only be enabled on one connector instance at a time.

MT:

Die Verzeichnissynchronisierung wird von der Konnektorkomponente des Diensts durchgeführt und kann jeweils nur auf einer Konnektorinstanz aktiviert werden.

Human MTPE:

Die Verzeichnissynchronisierung wird von der Connector-Komponente des Diensts durchgeführt und kann jeweils nur auf einer Connector-Instanz aktiviert werden.
Prediction PE vs. Actual PE
ML Model Inference Acceleration Solutions

Inference time comparison

CPU without acceleration:
400 ms/string

CPU + OpenVINO
130 ms/string, 3 x

GPU + TensorRT:
2.8 ms/string, 140 x
What’s next: APE (Automatic Post Editing)

MT Quality Prediction Service

Automatic Post Editing

TMS

Source

MT

MT Middleware

MT Engine 1

MT Engine 2

Other MT Engines
Q&A

Thank you