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

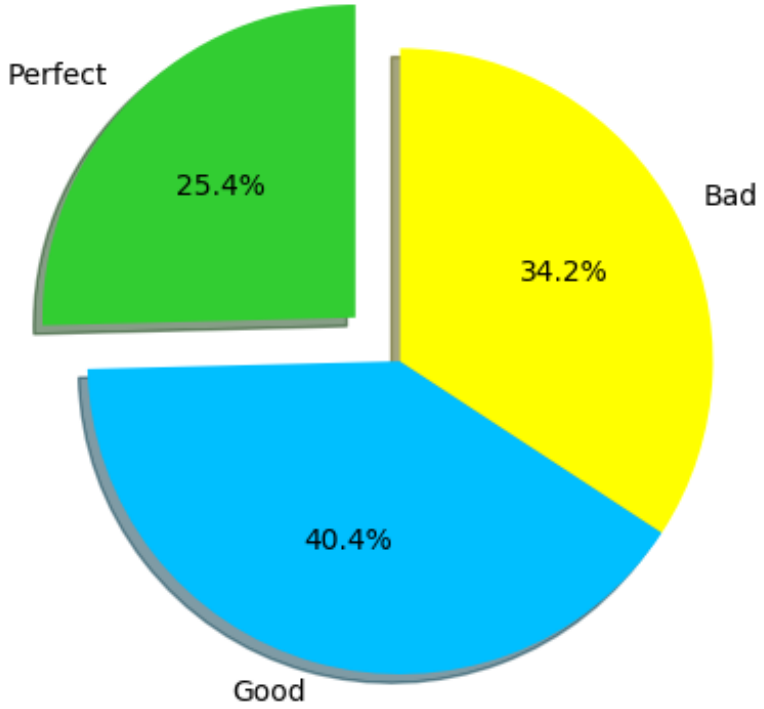
04 Inference acceleration

05 Future works

vmware

Program overview

Why prediction is needed



Perfect: PE% = 0 **Good**: 0 < PE% < 20% **Bad**: PE% > 20%

vmware

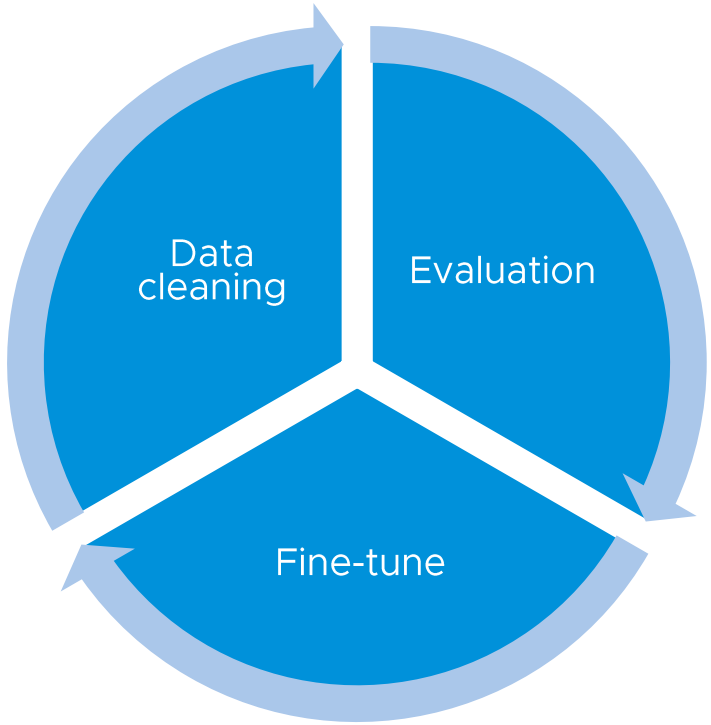
Program overview

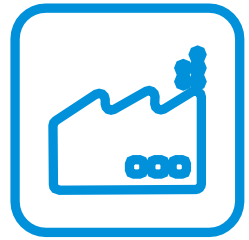
POC	Model fine- tune	Deploy to Stag.	Deploy to Prod.	Scenarios
• Data collection • Data washing • Conceptual design • Result analyze • Prove of concept	• Added source English as feature(English + MT as input, PE as the label) • Added regression to get linear output • Generated training data from fuzzy/ICE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

vmware

Data collection and model training

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



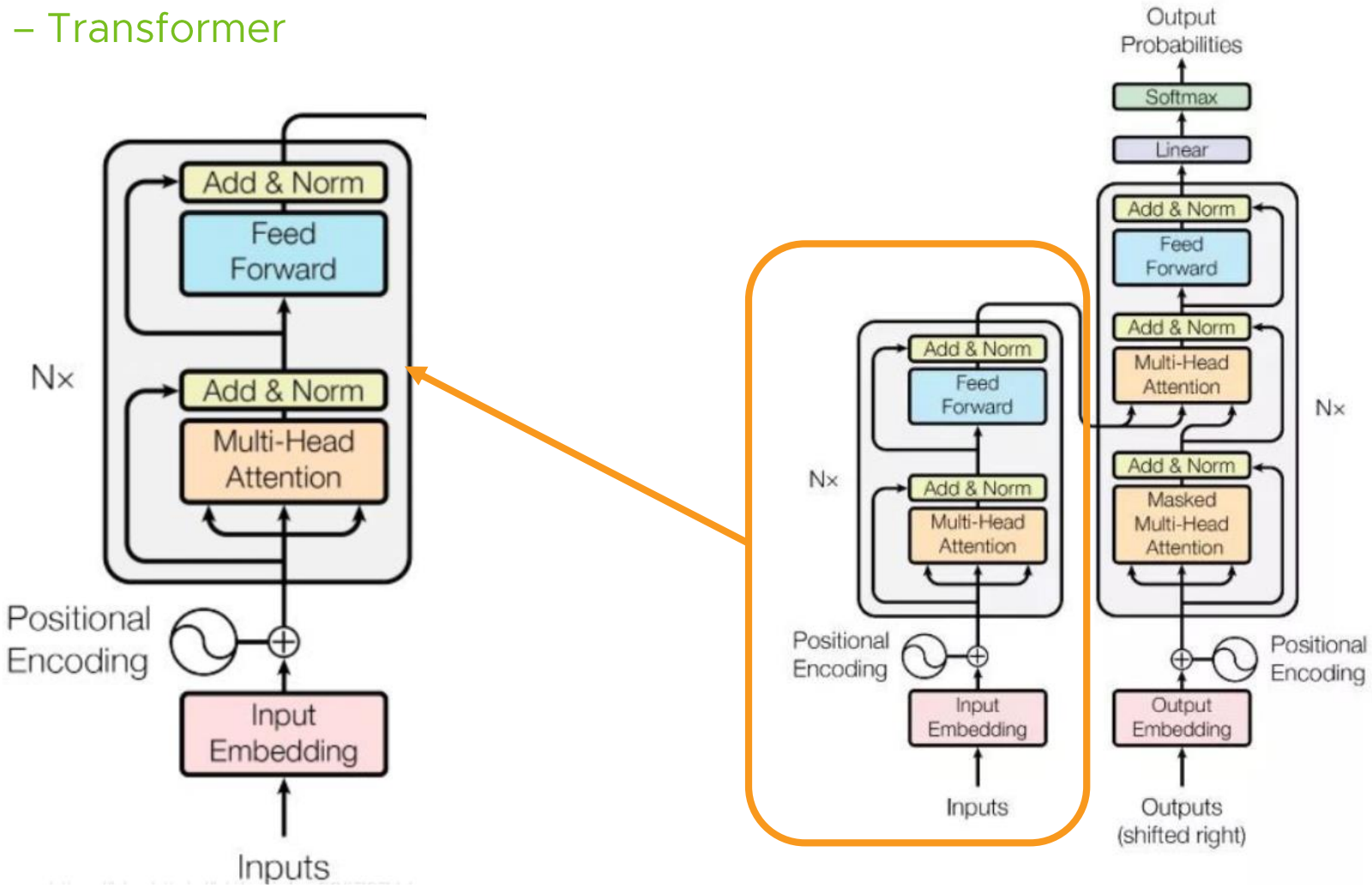


Integrate into production

vmware

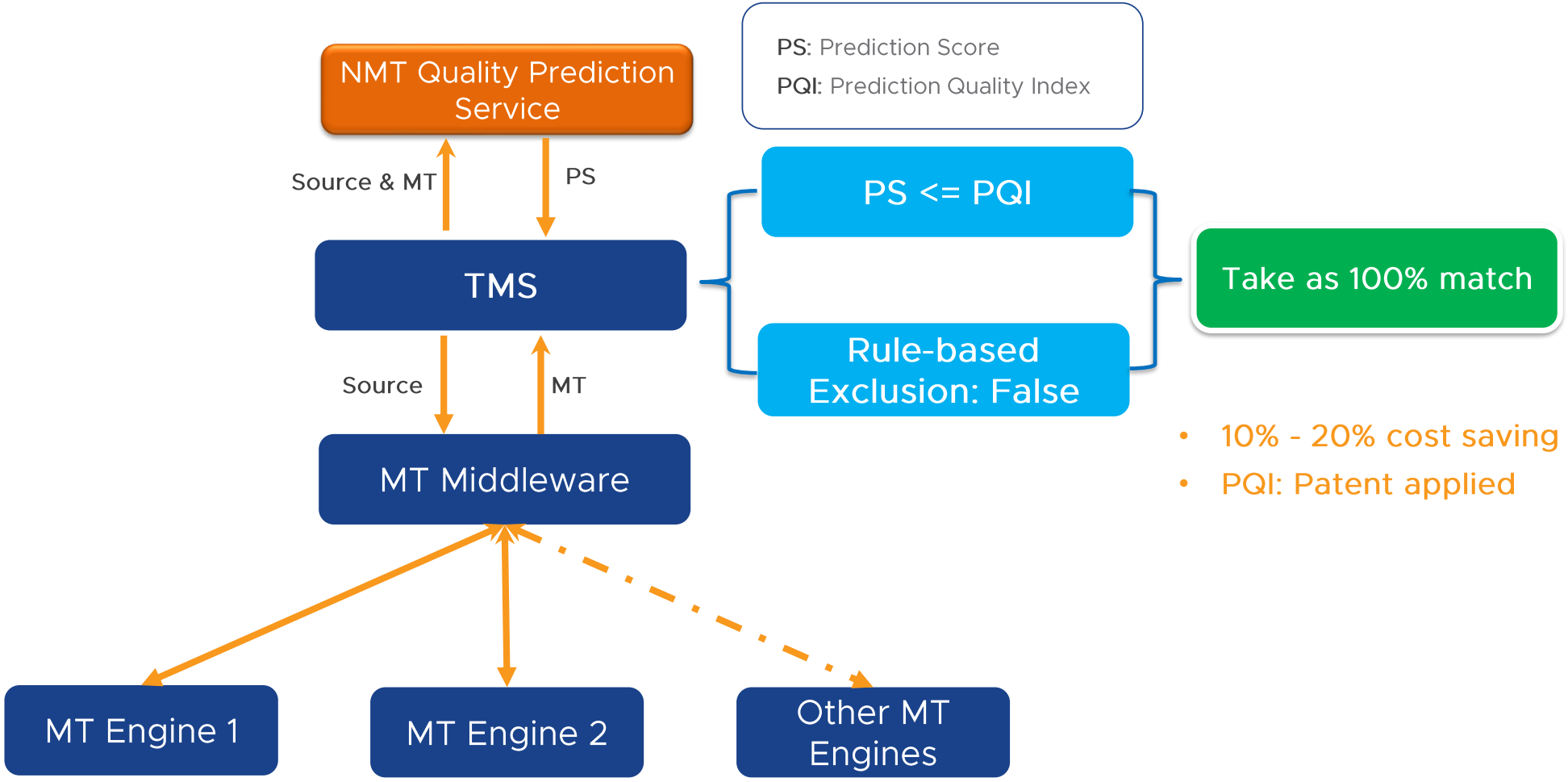
Data collection and model training

Algorithm – Transformer



vmware

Perfect MT scenario

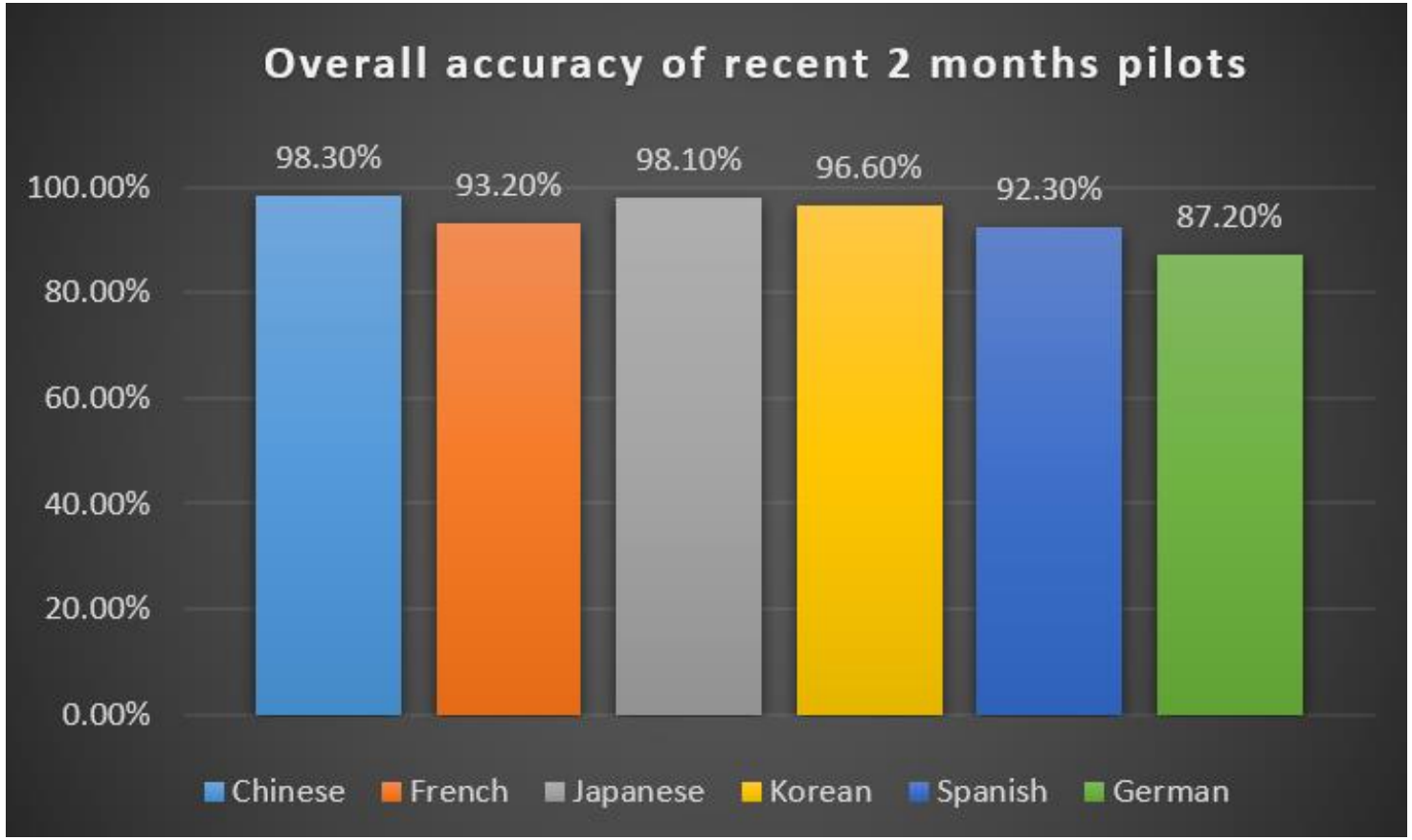


- 10% - 20% cost saving
- PQI: Patent applied

vmware

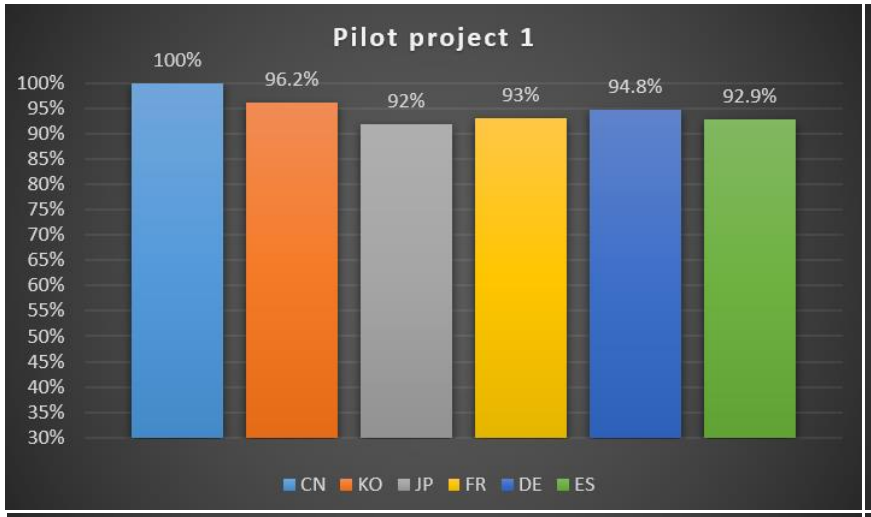
Perfect MT Scenario

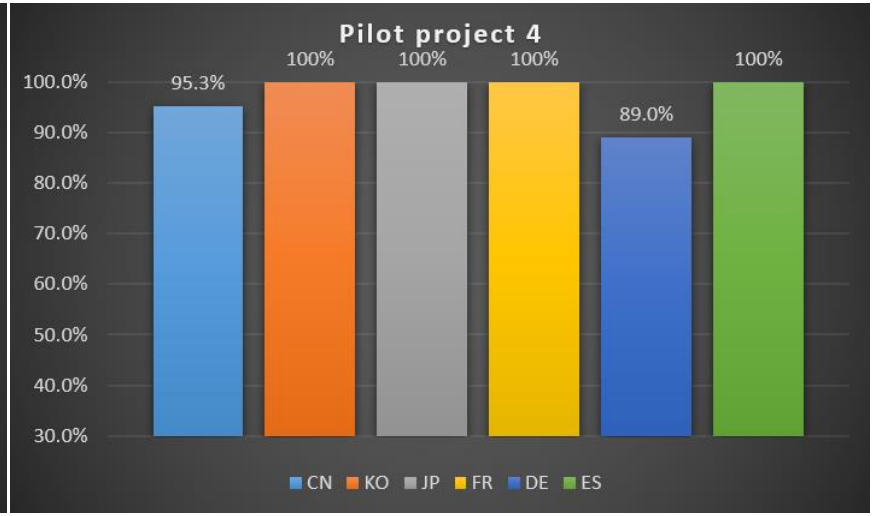
Overall accuracy



vmware

Perfect MT Scenario





vmware

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:

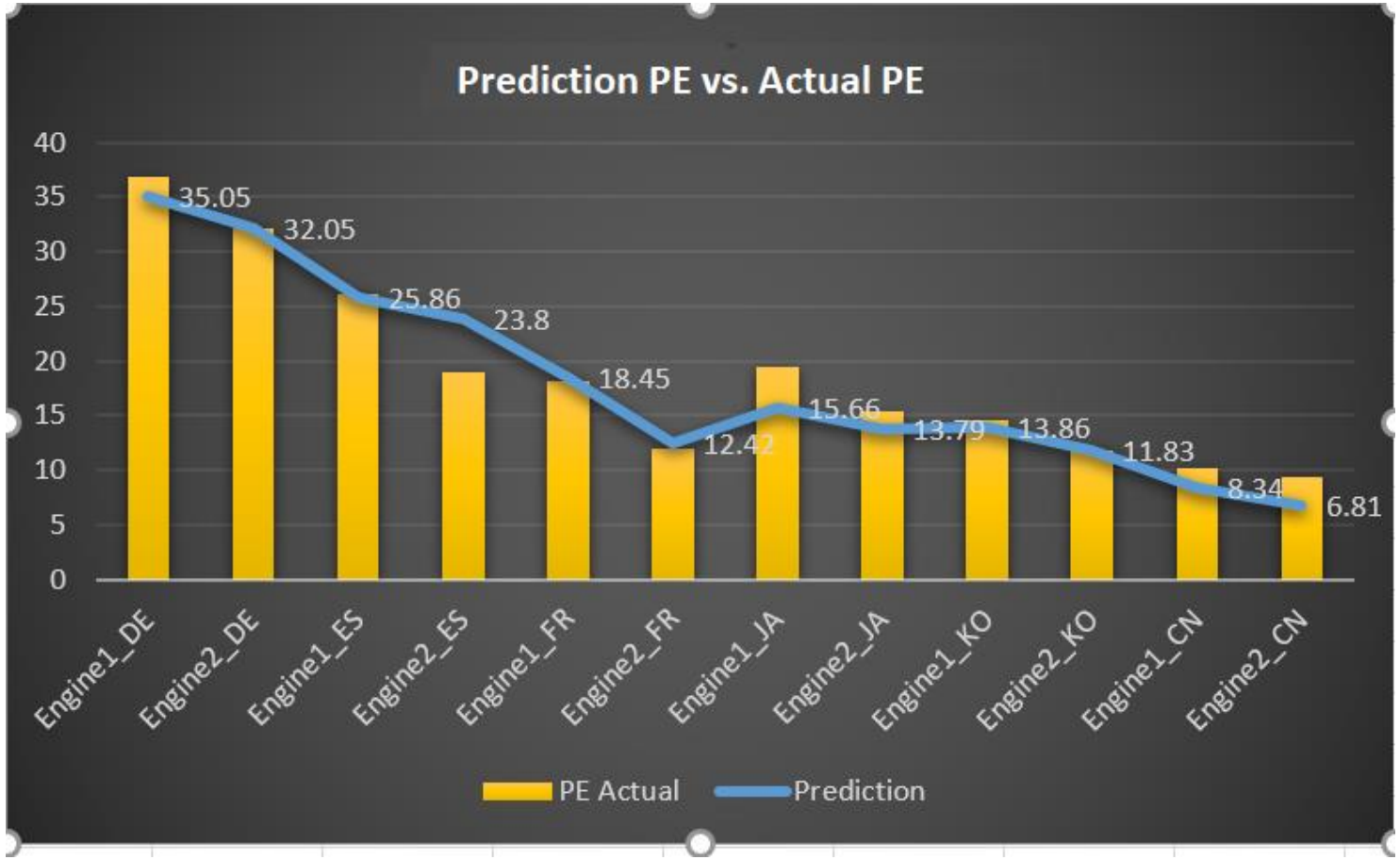
Die Verzeichnissynchronisierung wird von der Konnektorkomponente des -Diensts durchgeführt und kann jeweils nur auf einer Konnektorinstanz aktiviert werden.

Human MTPE:

Die Verzeichnissynchronisierung wird von der Connector-Komponente des Diensts durchgeführt und kann jeweils nur auf einer Connector-Instanz aktiviert werden.

vmware

Prediction PE vs. Actual PE



vmware

ML Model Inference Acceleration Solutions

Inference time comparison

CPU without acceleration:

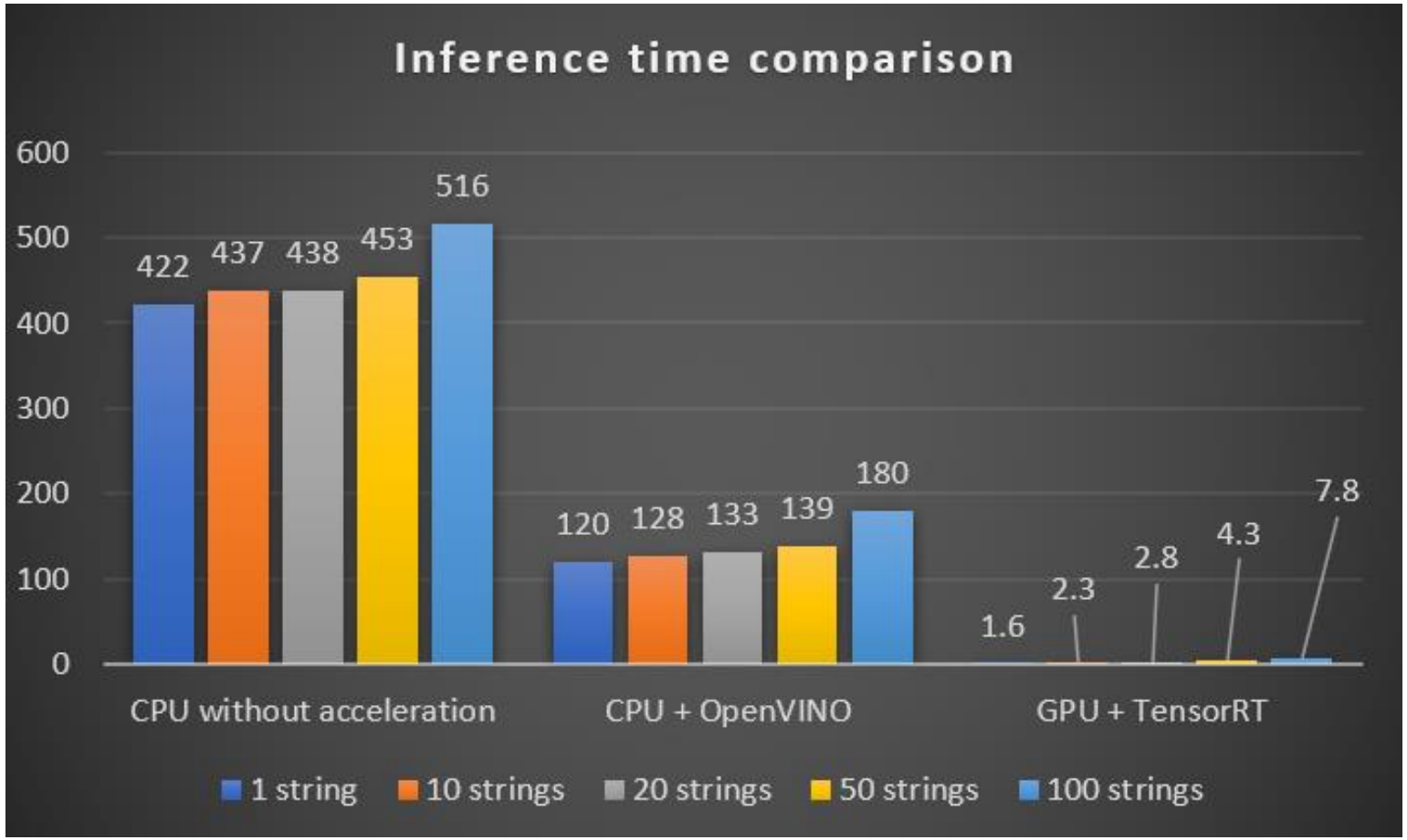
400 ms/string

CPU + OpenVINO

130 ms/string, 3 x

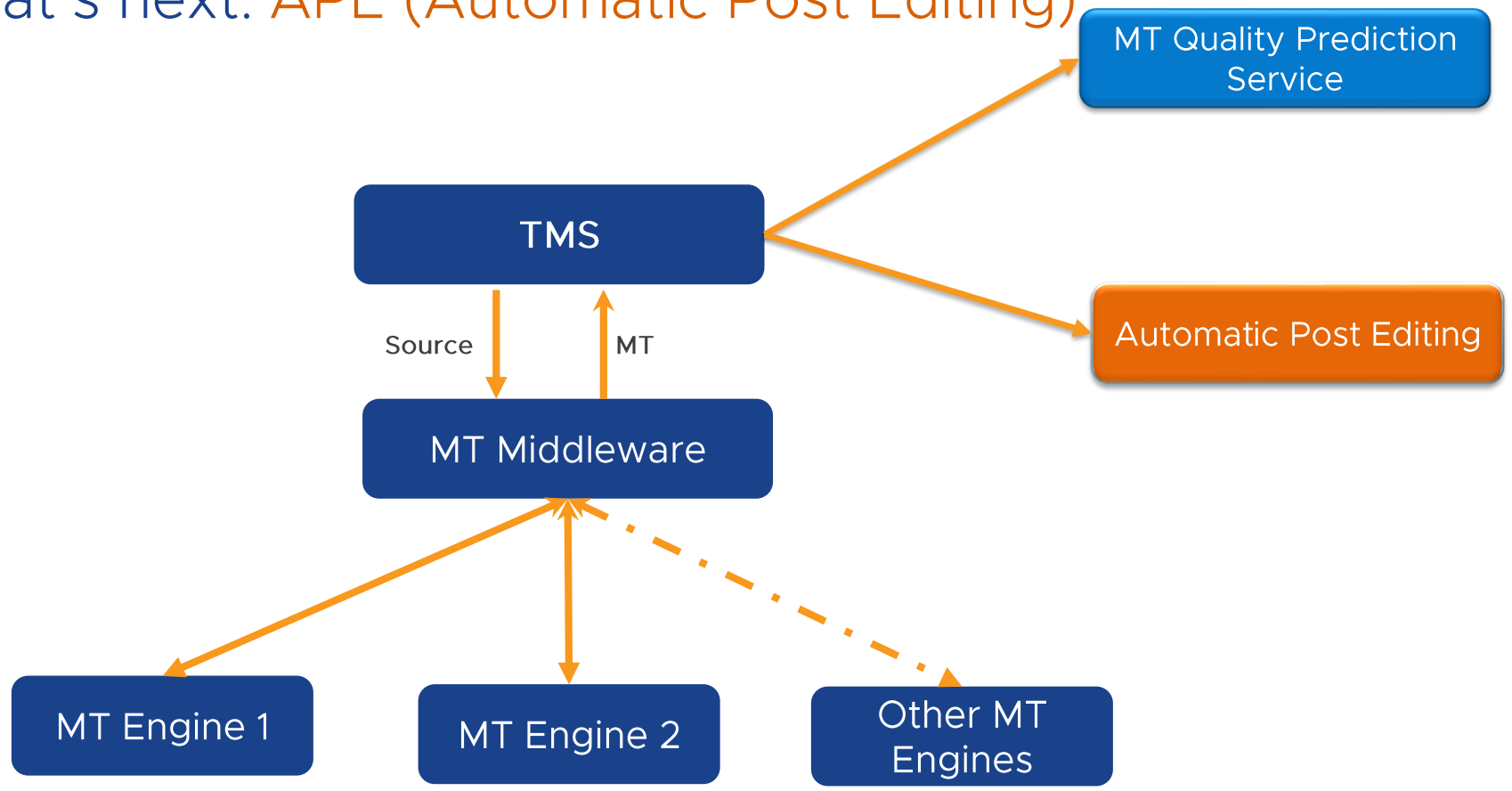
GPU + TensorRT:

2.8 ms/string, 140 x



vmware

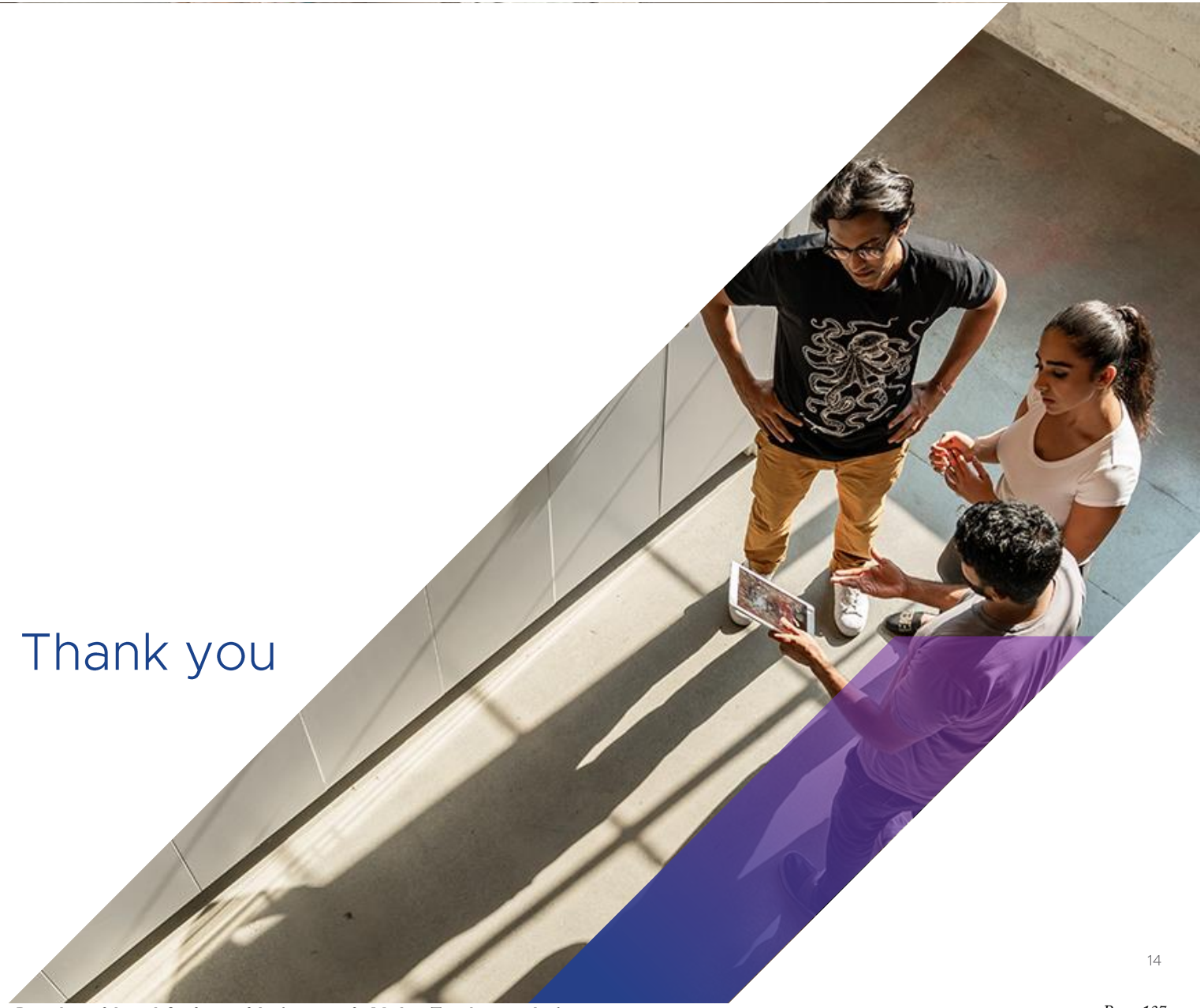
What's next: APE (Automatic Post Editing)



vmware

Q&A

Thank you



vmware

Proceedings of the 14th Conference of the Association for Machine Translation in the Americas
October 6 - 9, 2020, Volume 2: MT User Track