Data Analytics Meet Machine Translation

Allen Che and Martin Xiao

2022 Sep
$12.85B in revenue (USD) in FY2022

VMware is the leading provider of multi-cloud services for all apps, enabling digital innovation with enterprise control.

37,500+ employees

Allen Che
Senior Technical Localization Program Manager

Martin Xiao
Senior Technical Localization Program Manager
Objective

Business Manager

Localization PM

Technical Lead
Data is the key

**Discovery**
Discovery the data to uncover the pattern and trends.

**Insight**
Gain an Insight of the data to acquire the new value/new knowledge/new story.

**Predict**
Predict what could happen and make better and more scientific decisions.
Our Study – MT Quality

- **Perfect**: PE% = 0
- **Good**: 0 < PE% < 20%
- **Bad**: PE% > 20%
MT Solution – ML aided

- 3-in-1 MT solution
  - MT Customization
  - MT APE
  - MT Quality Prediction

- 20% quality improvement
- 10% - 20% cost saving
How the MT solution works

ML-Aided workflow

- MT Quality Prediction: auto quantify MT quality, ML based;
- MT Quality Prediction: Perfect MT use-case
- APE Service: auto post editing to improve MT quality, ML based
Challenge

- Monitor and alert
- Easy to search and manage data
- Real time updates with latest data
- Large data set, 7M bilingual strings every month
Solution – Elasticstack
Automate, Monitor and Visualize

Data Store
Search Engine
Analytics
Solution

Kibana
Explore, Visualize, Enage

Elasticsearch
Store, Search, Analyze

Integration
Connect, Collect, Alert

Visualize & Analyze
Explore data
Monitor and manage
Scalable
Real-time
Query & aggregations

Import the Machine Translation data

Beats
Logstash

Automate Data Collection
Monitor

Post-edit distance
Visualize the data
What’s Next
Intelligent Data Platform

- Automate the process
- API integration
- NLP layer on top of automation
What’s Next
Architect

APE Automation
MTQP Automation
MT Customization Automation
MT Quality Dashboard
TMS Monitor
MT Middleware
Quality Upstream

GPU servers
MT Engines

Data: Elastic Stack
TMS

Gitlab/CICD Pipelines

ESP/DECC/AWS
Live Q & A