

Scaling up automatic translation for software: reduction of post-editing volume with well- defined customer impact

Dag Schmidtke, Senior Program Manager
Microsoft E&D Global, Dublin
dags@microsoft.com

AMTA 2020



Automatic Translation for software (AT4SW)

Challenge

- Publish more MT for software without human review, with minimal customer impact
- MT quality is highly variable, both within and across languages

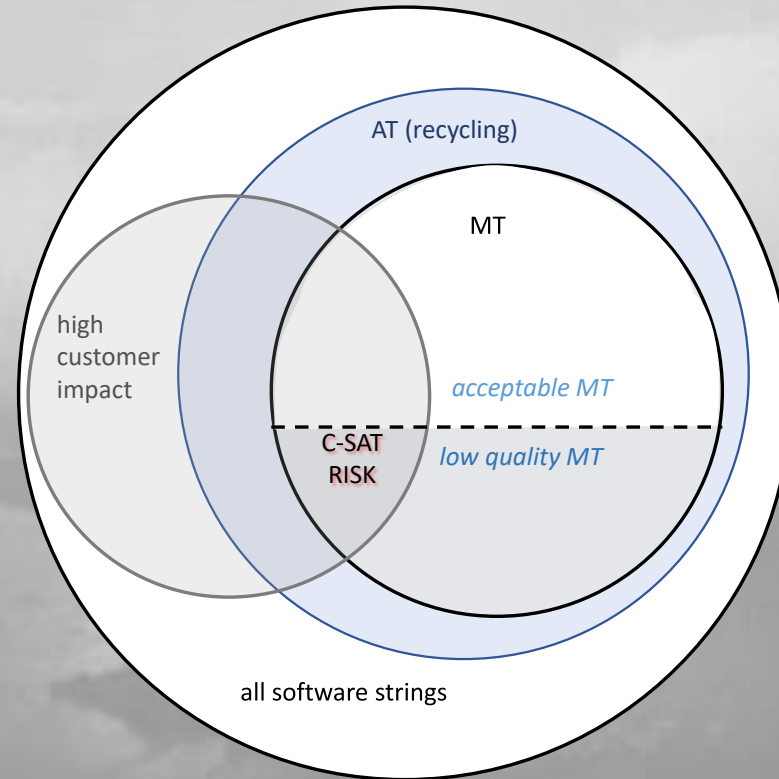
Approach

- Safe velocity: sw workflow with configurable constraints and quality gates
- Quality Estimation (QE) enables us to predict MT translation quality
- Workflow tuned to limit low quality MT to 10% of translation volume

Outcomes

- MT now used for 9% of published software translation volumes across 37 languages
- No notable negative impact on customer sat

Safe Velocity: managing risk

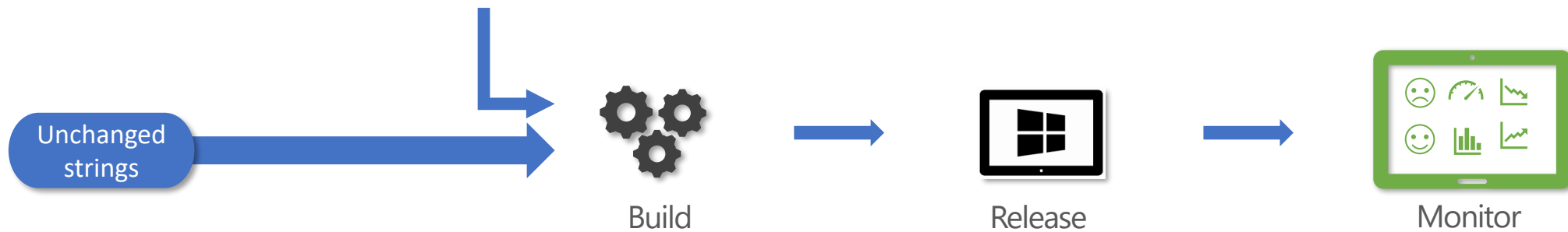
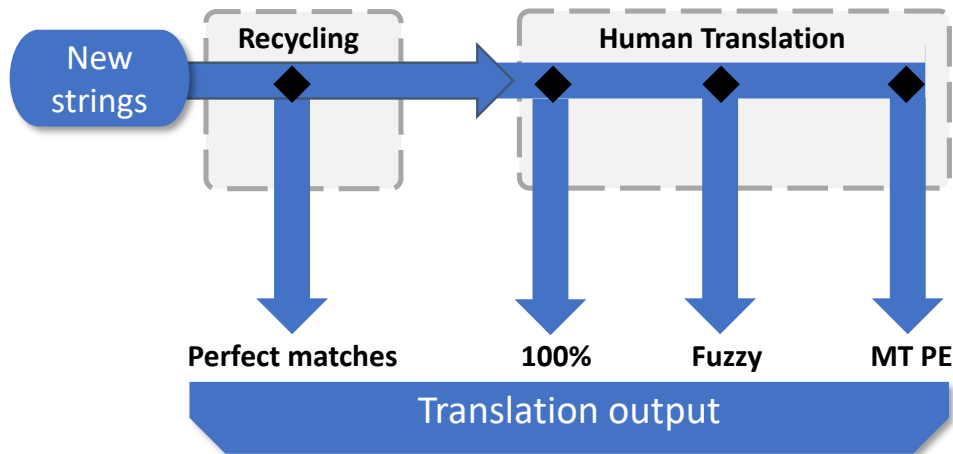


Automatic Translation (AT) for sw levers to maximize MT usage, with minimal SAT impact

- Improve MT and Optimize Quality Estimation (QE) to reduce low quality MT
- Protect high customer impact strings: exclusion, length thresholding
- Listen and respond to customer feedback

Software UI

Workflow



CORE TO USER EXPERIENCE



HIGH AMBIGUITY / MAJORITY OF SEGMENTS ARE SHORT (<5 WORDS)

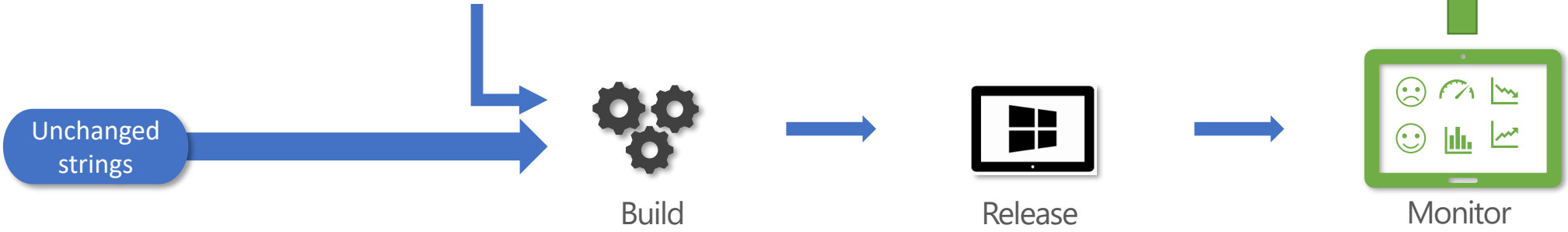
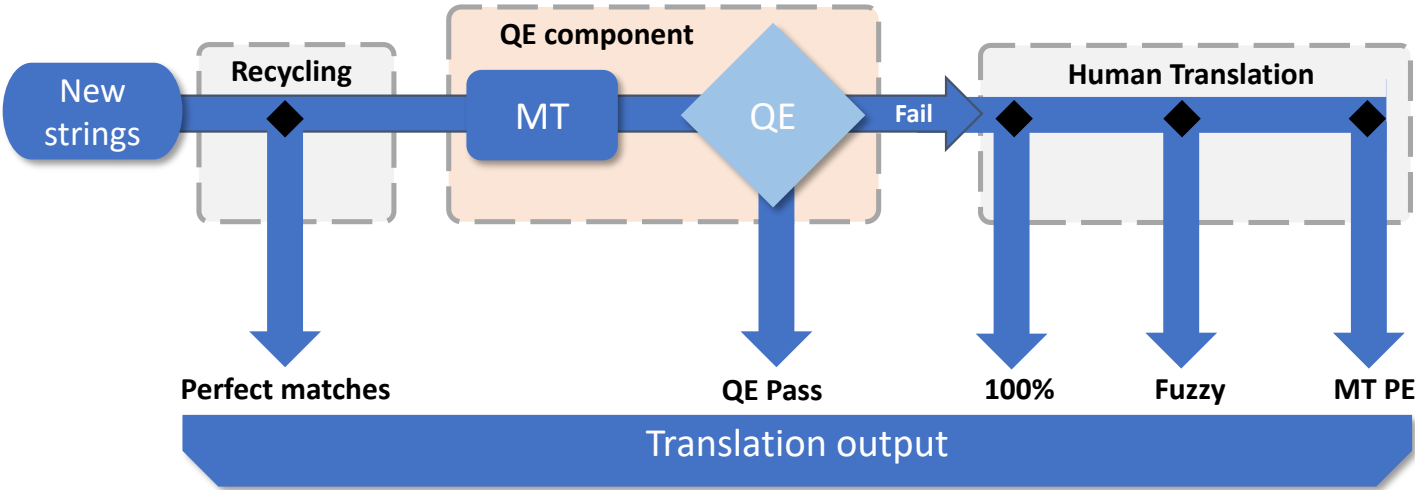




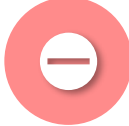
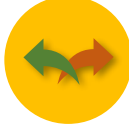

SPECIFIC TRANSLATION CONSTRAINTS (PLACEHOLDERS, COMPLEX PATTERNS...)

CAN WE DETECT WHEN MT IS GOOD ENOUGH AND DOES NOT REQUIRE POST EDITING?

Software UI

New workflow



-  ONE QE MODEL PER LANGUAGE
-  ESTIMATE THE TER SCORE
-  EXCLUDE PROBLEMATIC STRINGS
-  MODELS ARE NOT PERFECT
-  CALIBRATE FOR USER SATISFACTION

Exclusion for High Customer Impact



Why a need for exclusion?

- MT output quality can vary between string type/context & languages
- Some UI strings need get Human Review, as the risk of customer impact is high

Marketing	What's New	Legal
Welcome to Office Your place to create, communicate, collaborate, and get great work done.	“Starting from scratch is hard. QuickStarter automatically creates an outline for your topic of choice with suggested talking points and designs that make your presentation pop”	“By checking this box and entering your name below, you represent that you have read and understand above agreement, have authority to bind Customer, and that Customer agrees to be bound by the Agreement terms and the websites therein.”

Mechanisms for exclusion

- By resource: targeting specific words and phrases in strings, resource names, or developer comments
- By feature: not suitable or ready for MT, such as 'What's New', or resource groups with complex formatting

Initial target for exclusion: up to 20% of new words per month

Quality and customer impact: Error rate

We manage MT quality based on error rate: % of predicted low quality MT

- Based on volume of new words per product and month
- Assumption is users will tolerate a certain ratio of low-quality translations, without significant impact on customer satisfaction
- Historical human translation Linguistic Quality Assurance fail rate is 5%, by string
- MT error rate threshold, per product, language and month, is set to 10%, by word count – this is the amount of low-quality MT we tolerate

We use Quality Estimation (QE) to estimate the error rate

- Feature based ML model based on Quest++, trained on 100k+ segments /language
- MT low quality strings are those with a TER score >0.3, as predicted by QE
- QE threshold is calibrated per language, taking precision and throughput into account, against the 10% error rate

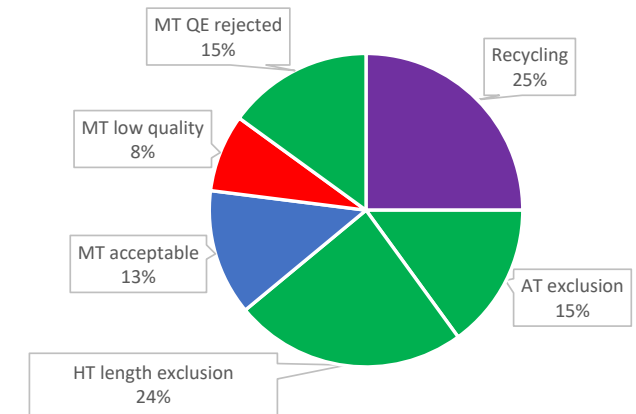
Calibration and MT error rate

Maximize AT volume against a MT error rate

- Recycle rate for contextual (perfect match) recycling
- High customer impact exclusion (AT exclusion)
- Length threshold for MT – we exclude short strings, <8 words
- QE precision and throughput per language
- This allows us to intentionally publish some low-quality MT

Example: QE threshold set to not exceed the error rate

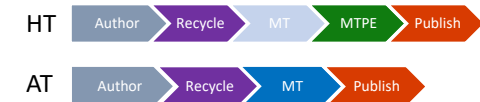
- We select the QE threshold for the right balance of throughput and precision, to hit the target error rate given volume in scope
- In the example, 36% of volume is in scope for MT. A QE threshold of 0.42 results in throughput of 58% and precision of 62%, and an 8% error rate



Green areas show HT workflow.

Purple AT workflow for recycling

Blue & Red AT workflow for MT



Scaling out AT4SW to a wider range of products

Goal for FY20: (Jul-19 to June-20) – expand AT4SW from Office to Windows products

Key Question: Would existing QE models provide sufficient accuracy, or need retraining?

- QE initially trained on Office product range, 2+ years worth of Post-edit data

Outcome: QE precision for Windows products sufficient to maintain MT volume level similar to Office products

- Good indication that our QE models are robust
- Office and Windows products are of a similar/overlapping domain

MT model training, evaluation, bug-fixing

- AT4SW makes use of Microsoft Translator custom models
- Automation and analytics in place to train and evaluate models for 90+ languages, for multiple domains
- Custom MT pre and post-processing in place for tag protection
- Custom training cleanup tools, aligned with pre-processing tools, to ensure we train on the same format text we process at runtime
- Monitoring of quality, analysis of post-editing, and collaboration with Translator team on bug-fixing

Development and optimization

- MT audit rate: measuring error rate in production
 - QE score assigned to all MTd strings, including those that get post-edited
 - Actual TER scores used to calculate Audit Rate: in production edit rate
 - Preliminary results indicate QE predicted scores and error rate is achieved in production
- AT4SW optimization to increase volume against error rate
 - Word count threshold reduction from 10 to 8 words in scope for MT QE
- Reduction of validation failures for MT by integrating upstream string information (dev comments) on placeholders

Summing up: 2 years on...

60_M

FLOW OF 60 MILLION WORDS PER YEAR
DISTRIBUTED ACROSS 37 LANGUAGES

0.4_M

VOLUME OF WORDS "QE
PASSED" EACH MONTH

9%

PROPORTION OF WORDS
SET TO "QE PASSED"

79%

CUSTOMER SATISFACTION
SCORES STABLE



NO SIGNIFICANT
INCREASE OF DSAT BY
LANGUAGE

Challenges



EXCLUSIONS &
TERMINOLOGY



PRODUCT
SPECIFIC TUNING



ACTIONABLE
FEEDBACK

References

- Glen Poor. 2018. Use more Machine Translation and Keep Your Customers Happy. Commercial Keynote at AMTA 2018, Boston
- Dag Schmidtke. 2016. MT Tresholding: Achieving a defined quality bar with a mix of human and machine translation. Paper presented at AMTA 2016 Users Track, Austin
- Dag Schmidtke, Declan Groves. 2020. [Automatic Translation for Software with Safe Velocity](#), in proceedings of Proceedings of Machine Translation Summit XVII
- Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level Translation Quality Prediction with QuEst++. In Proc. ACL, pages 115–120. Beijing, China

Q & A