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

Unchanged strings

Recycling

Human Translation

Translation output

Perfect matches

100%

Fuzzy

MT PE

Workflow:

- New strings
- Recycling
- Human Translation
- Translation output

Perfect matches, 100%, Fuzzy, MT PE

Workflow steps:

- Build
- Release
- Monitor

Can we detect when MT is good enough and does not require post editing?

Core to user experience

High ambiguity / majority of segments are short (<5 words)

Specific translation constraints (placeholders, complex patterns...)

Unchanged strings

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

Page 113
Software UI

New workflow

- New strings
- QE component
- MT
- QE
- Human Translation
- Perfect matches
- QE Pass
- 100%
- Fuzzy
- MT PE
- Translation output
- Unchanged strings
- Build
- Release
- Monitor

- 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

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
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…

FLOW OF 60 MILLION WORDS PER YEAR DISTRIBUTED ACROSS 37 LANGUAGES

<table>
<thead>
<tr>
<th>0.4M</th>
<th>9%</th>
<th>79%</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOLUME OF WORDS “QE PASSED” EACH MONTH</td>
<td>PROPORTION OF WORDS SET TO “QE PASSED”</td>
<td>CUSTOMER SATISFACTION SCORES STABLE</td>
</tr>
</tbody>
</table>

NO SIGNIFICANT INCREASE OF DSAT BY LANGUAGE

Challenges

EXCLUSIONS & TERMINOLOGY

PRODUCT SPECIFIC TUNING

ACTIONABLE FEEDBACK

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


• 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