Office

MT Thresholding: Achieving a defined quality bar with a mix of human and machine translation

Dag Schmidtke
Senior Program Manager, Office Global Services & Experiences
Microsoft Ireland

Microsoft

Recycling & Machine Translation in Office

Goal: Maximise use of Recycling and Machine Translation, while protecting Customer Satisfaction

- Focus spend: Human translate high priority and most popular content
- Recycle as much as possible and machine translate the rest, publish and upgrade based on quality and traffic
- Increase velocity & reach with added coverage

Recycling

Reuse of existing high quality translations

Automated in production process

Typically reduces wordcount & cost by 60 to 70%

Human Translation with MT Post-Editing (MTPE)

Improve MT output with human translators

No quality degradation

Applied after recycling

Part of production process for UA and UI

In use for 35+ languages

MT Publishing

Machine translation published without human editing (raw-MT)

Applied after recycling

Used for long tail content, speed

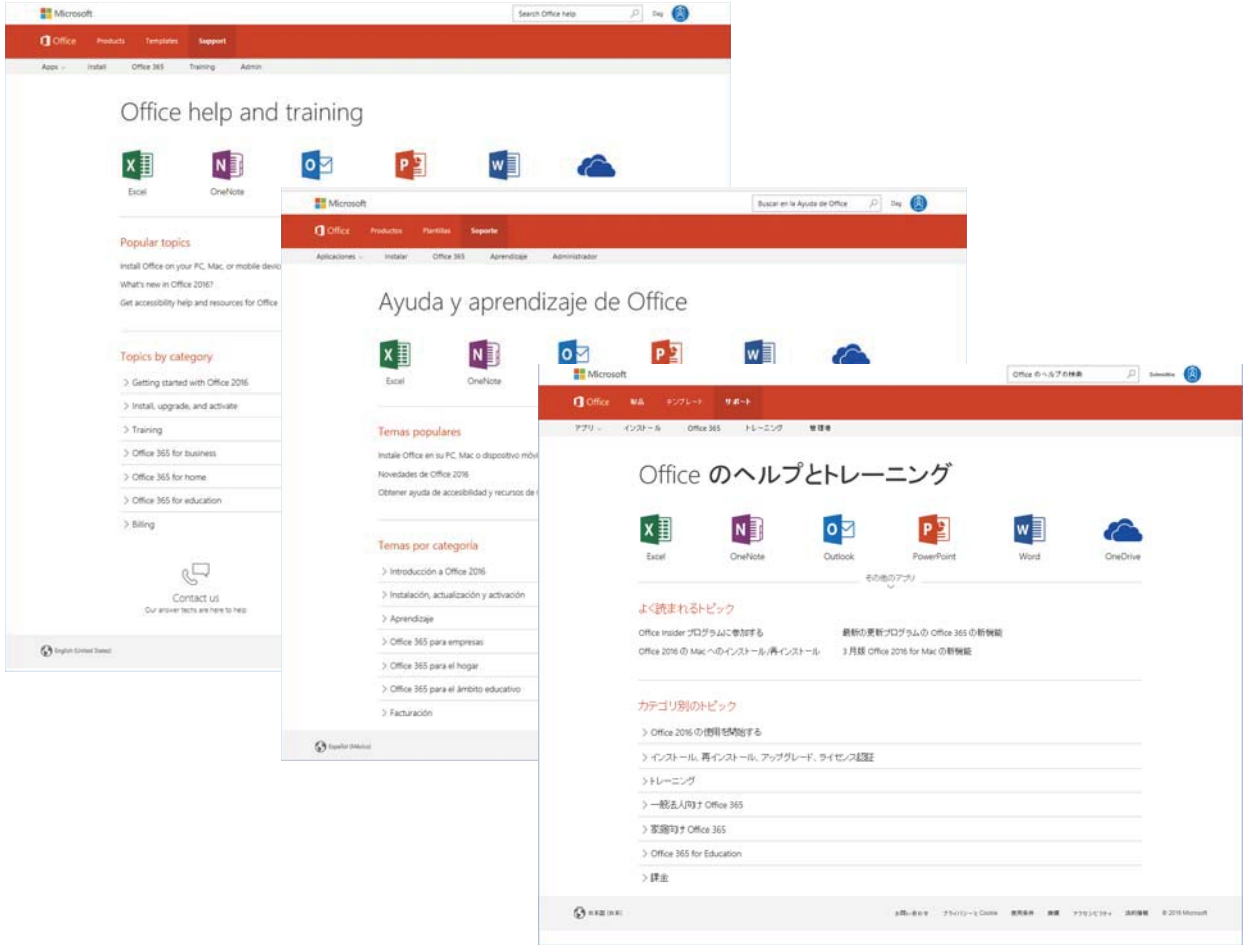
In use for 38 languages

Human Translation (HT) Workflow

MT Publishing Workflow

Microsoft

Use-case: MT Publishing for Office Help



Support.Office.COM (SOC)

- End-user help & training
- Large scope: 40 languages x 15k articles
- About 2.9 bn PVs /year, 45% non-English
- Significant translation effort
- Can MT help?

Microsoft

Challenges

- Is MT good enough for end-user help?
  - MT quality unpredictable and hard to measure
  - What is the right metric for 'MT quality'?
  - Will the Office end user audience accept MT?
- How to prioritize human versus machine translation?
- How to achieve scale?
  - Office Translation requirement: 100s of millions of words/year
- How to listen to customers and respond?

Microsoft

MT Publishing: Our Approach

1. Plan: Establish KPIs
   - Quality bar: 'Acceptable'
   - Speed: <24 hours
   - Scope: Low PV topics

2. Prepare: Engineering
   - Benchmark MT quality
   - Automation in platform
   - Internal telemetry
   - Business Intelligence: traffic and ratings

3. Deploy: Optimised MT
   - Custom MT domains with Microsoft Translator Hub
   - Recycling: re-use of high quality translations
   - Quality gating with thresholding

4. Iterate and adjust
   - Active monitoring of usage and ratings
   - Traffic-based Upgrades
   - Adjust thresholds
   - Increase MT scope

Microsoft

Quality model

Quality bar – 'Acceptable'

- MT Publishing needs to reach a minimum bar to be usable
- Starting metric: 2.5 /4 for human evaluation
- Ongoing metric: within 10% of Human Translation User Rating (CSAT), for each language

Thresholding - based on initial human evaluation and recycle rate per article

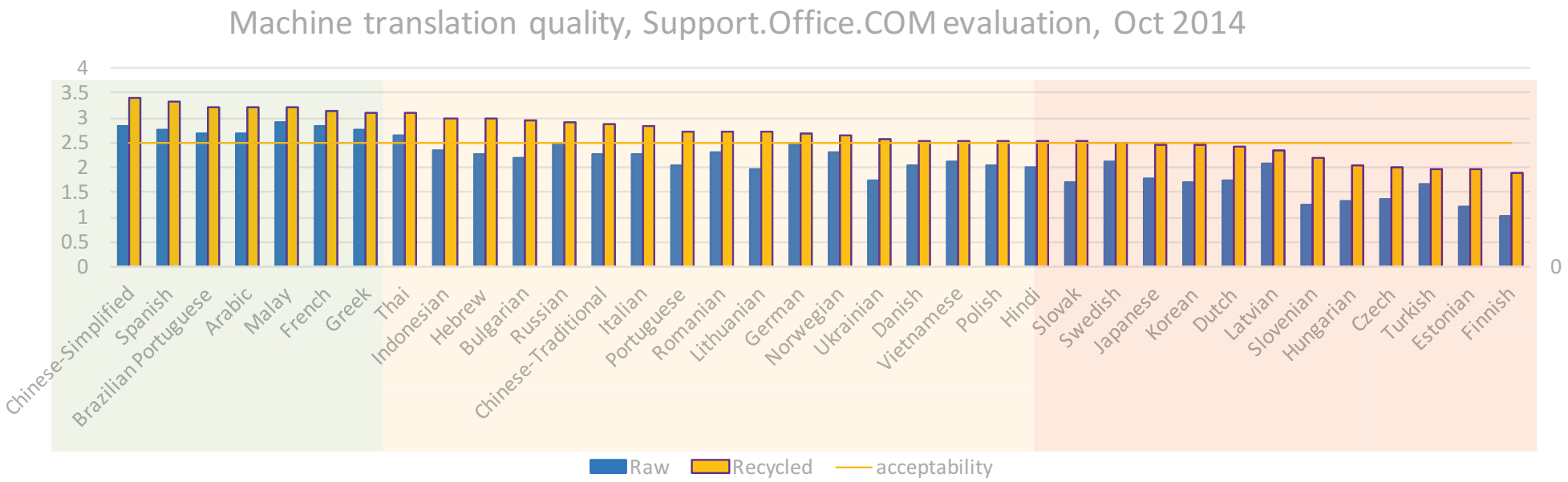
- Good quality MT (>=2.5/4): article published without restriction
- Medium quality MT (>=2.5 with recycling): article recycle rate of >=50% needed to publish
- Lower Quality MT (<2.5 with recycling): article recycle rate of >=80% needed to publish

Iterate and adjust

- High traffic MT assets (within top 70%) upgraded to HT, to ensure optimal customer experience
- HT used for high priority articles based on source meta-data

Microsoft

Initial MT Human Quality Evaluation



Methodology

- Human evaluation, 3 reviewers per language
- Judged on scale of 1-4, with 2.5 set as acceptability threshold for production use
- 10 help articles x 36 languages: 5 with 50% recycling, 5 with low or no recycling

Results: Variable MT quality

- 8 languages have good enough, 'acceptable' MT quality
- 16 additional languages reach quality bar only with use of recycling, medium quality
- 12 final languages have lower quality, did not meet the quality bar even with recycling

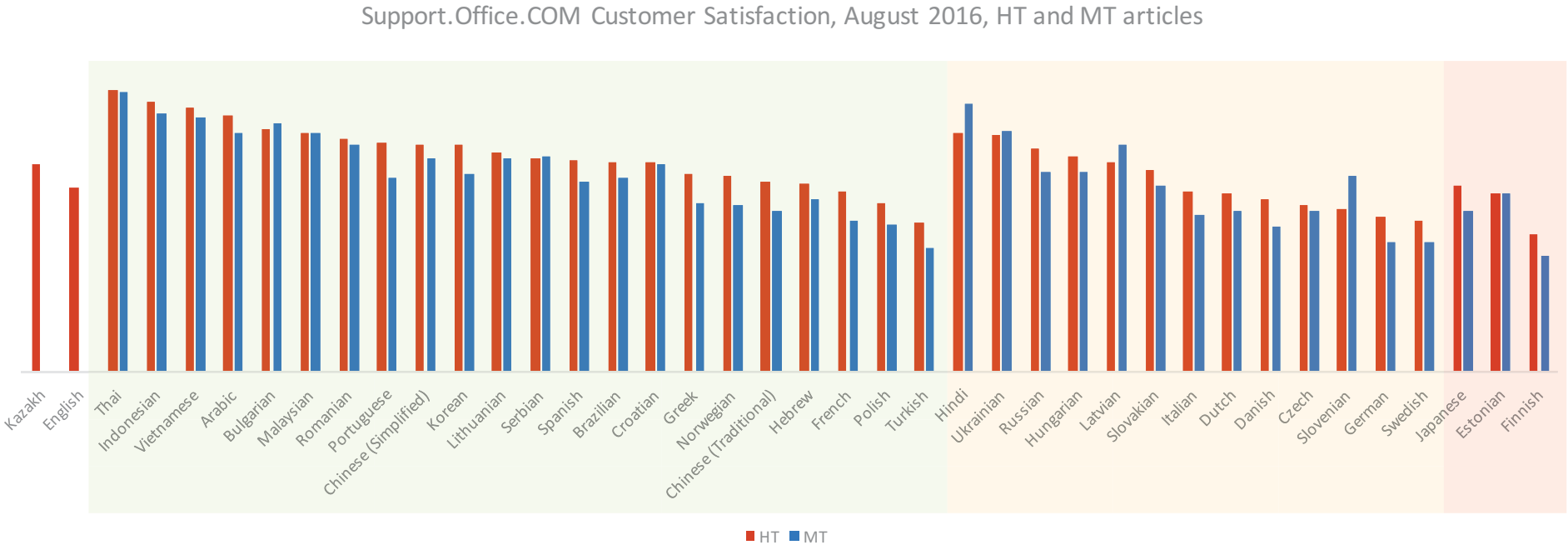
Microsoft

Progress

- June 2015
  - MT Publishing in use for 38 languages, 20% of monthly translation volume
  - Initial thresholding: 8 languages with 0 threshold
- November 2015
  - MT Publishing used for >50% of monthly volumes
  - Thresholding adjusted: 18 languages with 0 threshold, 15 with 50% threshold
- August 2016
  - MT Publishing used for >70% of monthly volumes
  - 47% of live articles published through MT pipe, generating 15% of traffic
  - Thresholding adjusted: 22 languages with 0 threshold, 13 with 50% threshold

Microsoft

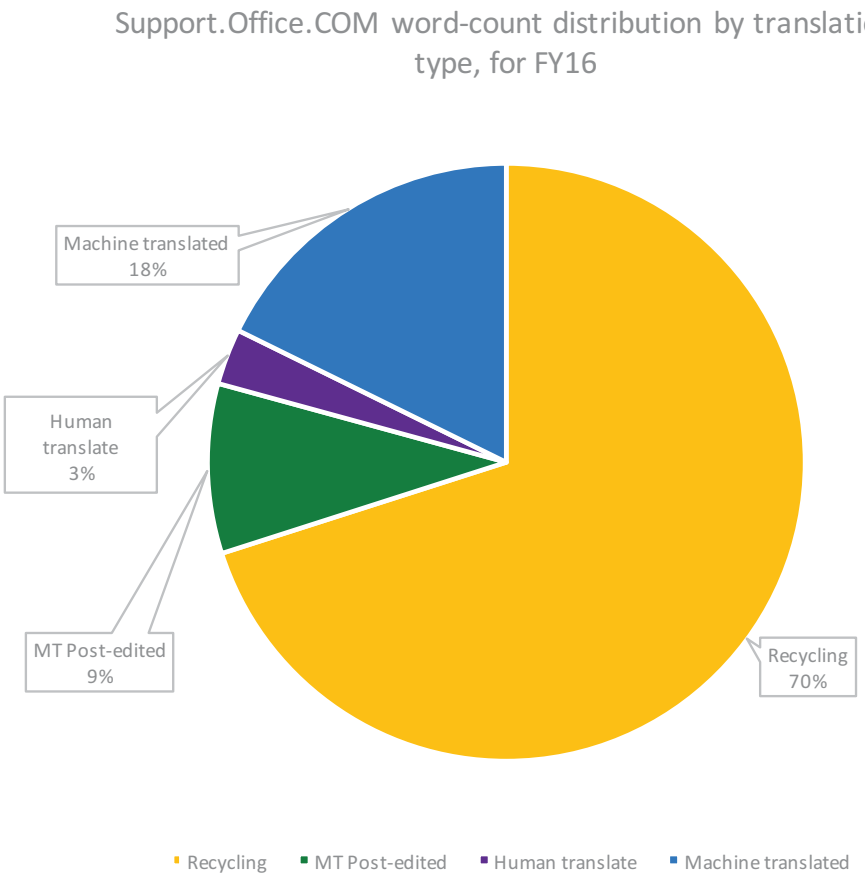
Support.Office.COM Customer Satisfaction



- CSAT based on Article ratings, question: '*Was this information helpful to you?*' (yes/no)
- Grouped by thresholding level: none vs 50% vs 80% recycle rate
- MT within 10% of HT for all languages, except Portuguese (11%)

Microsoft

Summary, Lessons & What's next



- MT Publishing can be used at scale, for end-user content
- Recycling helps extend the scope for lower quality MT languages
- Thresholding lets us control unpredictability in MT quality
- User acceptance of MT is greater than offline MT evaluations suggest

Still to do

- Some languages still need work

Coming

- Neural MT
- Experiment: MT Publishing for

Microsoft