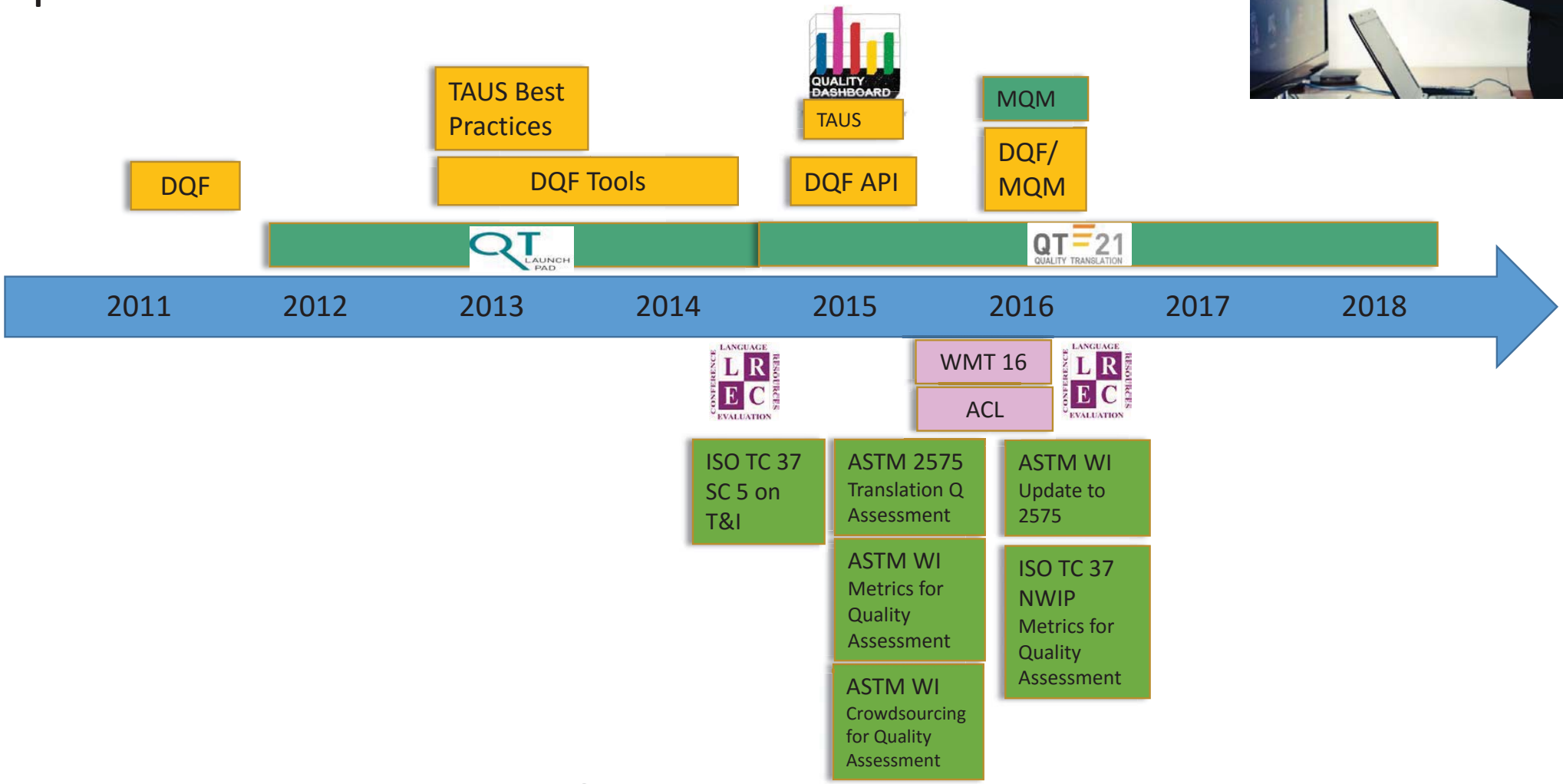
Assessing Translation Quality Metrics

29 October 2016

Jennifer DeCamp

What is going on with assessing quality of production translation?



Why is this important to us?

- These standards could have a real impact
  - Contracting organizations
    - LPTA (Lowest Price/Technically Acceptable)
    - Best Value
  - Potential for being used as policy
  - We use tools from TAUS integrators

- But TAUS, MQM, and ASTM based on industry surveys, requirements, practices, and efforts (e.g., TAUS, GALA, SAE)
  - Need government requirements and review

What do we do?

- Build our awareness of what is happening in this space
  - Understand marketing vs. reality
  - Assess impact
- Define our user requirements and make those requirements known in ASTM and ISO, and to companies implementing TAUS software
- Find ways we can help
  - Definitions of metrics, measures, etc. compatible with current systems
  - Definitions of specific metrics and error types
  - Contacts with the research community and their extensive findings
  - Coordination between standards efforts (e.g., AMTA workshop)
  - Close review and possible testing of tools and standards

What are we doing in this presentation?

- Raising awareness
  - What is the state of assessment production translation quality?
  - What are these standards and efforts?
  - What are other approaches?
  - What is a first cut at decisions?
  - What was said in the workshop on October 28?
- Discussing next steps

What is the state of approaches to MT or professional translation evaluation?

"Current approaches to Machine Translation (MT) or professional translation evaluation, both automatic and manual, are characterized by a high degree of fragmentation, heterogeneity, and a lack of interoperability between methods. As a consequence, it is difficult to reproduce, interpret, and compare evaluation results."

Rehm, G., A. Burchardt, O. Boja, C. Dugast, M. Federico, J. van Genabith, B. Haddow, J. Hajič, K. Harris, P. Koehn, M. Negri, M. Popel, L. Specia, M. Turchi, and H. Uszkoreit, (2016). Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem. In *Language Resources Evaluation Conference Proceedings*. Retrieved June, 2016, from http://lrec2016.lrec-conf.org/en/about/conference-proceedings/

How well do these measures apply to human-translated material?

"Quality measured by BLEU, NIST, METEOR etc. does not indicate the type of quality problems," and that "these metrics are also better suited for measuring progress in the 'ugly' or 'bad' sectors of the quality spectrum….Even the human evaluations usually by ranking, often done by CS researchers and students, do not help the human translators….and the LISA [Quality Assessment] QA model, EN-15038 and current International Organization for Standardization (ISO) work on a successor, are not known and not used in MT research."

Uszkoreit, H. and A. Lommel. (2014). Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment. Retrieved October, 2016 from http://www.qt21.eu/launchpad/sites/default/files/MQM.pdf

What are some of the problems?

"Humans differ in their understanding of quality problems, their causes, and the way to fix them."

Factors impacting human identification and classification of errors, include:

- Disagreement as to the precise spans that contain an error
- Errors whose categorization is unclear or ambiguous
- Differences of opinion about whether something is or is not an error or how severe it is."

Lommel, A., M. Popović, A. Burchardt (2014a). Assessing Inter-Annotator Agreement for Translation Error Annotation. Workshop on Automatic and Manual Metrics for Operational Translation Evaluation. *Proceedings of the Language Resources Evaluation Conference*, 2014. Retrieved June, 2016 from http://www.lrec-conf.org/proceedings/lrec2014/index.htm

What is standard practice in industry?

- Methods for assessing production translation quality
  - Extent to which the product met customer specifications
  - BLEU scores
- Demand for more consistent means of conducting assessments

O'Brien, S. (2012). Towards a Dynamic Quality Evaluation Model for Translation. *The Journal of Specialized Translation*, 17:January, pp. 55-77.

What is industry doing? TAUS

- Translation Automation User Society
- Dynamic Quality Framework (DQF), 2011
- DQF tools, 2013-2014
- Quality Dashboard, 2015
  - Some capabilities available to anyone; others available only to members
  - Developed and copyrighted by TAUS
- DQF API and member integrators, 2015
- Also have a Productivity Dashboard

What else is industry doing? MQM

- Multidimensional Quality Metrics
- Developed and copyrighted by DFKI and QT LaunchPad
- Based on the following definition by Melby
  - "A quality translation (1) demonstrates required accuracy and fluency (2) for the audience and purpose and (3) complies with all other negotiated specifications, taking into account end-user needs"
- Provides
  - A hierarchical catalog of issue types
  - Dimensions (based on ISO/TS-11669) to guide users in selecting appropriate issue types
  - A method for declaring/describing a particular metric
  - An inline format for tagging issues in XML files
  - A reporting format with scoring formula for determining scores/acceptance
- Error typology integrated with the one from DQF
- Study indicated "*Low inter-rater reliability but better with classifying errors using MQM than with identifying errors*" (Snow 2015).
- Basis for ASTM WK 46397 *Language Quality Assurance*

What are these standards?

F15.48 Committee on Language Services and Products/
Subcommittee on Language Translation

ASTM F2575 *Standard Guide for Quality Assurance in Translation*	2014
ASTM Work Item (WK) 47362 *Standard Practice for Quality Assurance in Translation*	2015
ASTM WK 46397 *Language Quality Assurance*	2016
ASTM WK 46396 *New Practice for the Development of Translation Quality Metrics*	2016

Technical Committee 37 on Terminology and Other Language and Content Resources/
Subcommittee on Translation, Interpreting and Related Technology

ISO/NP 21999 *Translation Quality Assurance and Assessment – Models and Metrics*	2016

What are options for assessing production translation quality?

Customer Requirements

- ✓ Consistent with Quality standards
- ✓ Documents requirements
- ✓ Enables education
- X Often still need "accuracy" measure
- X Hard to use across community
- X Requires extensive customer education

E.g. ASTM F2575, ASTM WK 46397, MQM

Confidence Prediction

- ✓ Fairly good for highly repetitive work
- X Less useful when translation is less repetitive with fewer terminologies or translation memories
- X Predicts amount of post-editing needed

E.g. Xerox and SDL Confidence Measures for raw MT

Capability, Proficiency

- ✓ Promotes process improvement
- X Time consuming
- X Info not always available
- X Difficult for comparisons
- X Poor skill measurements

E.g. ASTM 2575; productivity measures; skill measures

Adequacy, Fluency

- ✓ Focus on deliverable
- ✓ Consistent with government translation and acquisition practices
- ✓ Removes variability from process, people, tools
- ✓ Applicable for material with unknown process/people
- ✓ Useful for comparisons
- X Labor intensive
- X Slow
- X Low interrater reliability

E.g., ATA Certification scoring; automated measures; crowdsourcing and other quality estimation

Quality, Successful Task Completion

- ✓ Consistent with Quality standards
- ✓ Shows real world value
- ✓ Can show threshold
- X Rarely independent
- X Often outside control of translation group
- X Labor intensive
- X Often a time lag

E.g. Army Operational Testing via Task Based Assessment; chart completion; responses to "Was this site useful?"; number of visas processed

What is a first pass at decisions needs?

Population	Decision
Customer	Is the translation ready for use?
	Does additional work needs to be negotiated?
	Should the provider be paid?
	Did the translator or LSC provide good value?
	Should pricing be adjusted for future contracts?
Translator or LSC	Is the translation is ready for delivery to customer?
	If not, what else needs to be done?
	Does additional work needs to be negotiated?
	Should pricing be adjusted for future contracts?
	How do current practices and workflows compare to proposed ones?
	How do tools compare?
	How does the translator or the company compare to others?
LSC	Are the specific translators doing a good job in this language pair, domain, etc.?
	Is translator performance being affected by stress, fatigue, or other factors?
	Should the translators receive pay increases?
	Are some translators better than others at certain types of work?
End user	How reliable is the translated information?
	Is a re-translation warranted?
Researchers & Developers	How can we improve the tools?
	Does one tool or process work better than another?
	Can we improve the translation (e.g., through annotated data sets)?

What are government decisions needs?

What happened at the AMTA Workshop on Assessing Production Translation Quality?

- Facilitator
  - Jennifer DeCamp Chair, ATA Standards Committee; member ASTM, ISO, ILR
- Industry
  - TAUS
    - Achim Ruopp TAUS Director of Research and Development
  - SDL
    - Daniel Brockman SDL Director of Product Management
  - MQM
    - Alan Melby Co-Author of MQM
- Standards Groups
  - ASTM
    - Amanda Curry Chair, Translation Subcommittee
  - ISO
    - Sue Ellen Wright Head of U.S. Delegation; member ASTM
    - Monika Popiolek Chair, WG on ISO/NP 21999
  - Interagency Language Roundtable (ILR)
    - Maria Brau Chair, Translation Subcommittee

Summary

- Different populations with different decisions, decision factors, and metrics
- Inconsistent practice
- Efforts that are “fragmented, heterogeneous, and non-interoperable” (Rehm et. al., 2016)
- Conflicting terminology
- Still a divide between the MT research and HT practitioners
- Promising efforts to improve evaluation of production translation quality (TAUS Quality Dashboard, MQM, standards)
  - Extraordinary outreach
  - But based on industry practice, surveys, requirements, and needs

Recommendations

- Negotiate and document a common terminology and structure for evaluation (e.g., metrics, measures, methods, tools) and for the specific metrics (e.g., adequacy, fluency) and error types.
- Analyze the relationship between decision needs, metrics, and measures.
  - Do the metrics (e.g., fluency and adequacy) meet the needs of the decision-maker?
  - What do specific measures actually say about the metrics, and how may that be presenting an inaccurate or incomplete picture for the decision-maker?
- Leverage and/or conduct analysis on how standards, methods, measures, and tools support specific translation requirements.
- Provide resources for language service providers and others to easily access information on relevant metrics and measures.
- Provide standards within a framework, and reference that framework at the beginning of each standard along with the audience of the standard and the part of the problem that it is addressing.
- Develop more best practices documents and other guidelines for developing tools; encourage adoption through funding organizations and through professional organizations.
- Continue to bring together the research and language services communities.
- Increase outreach from the standards committees to researchers, decision-makers, and other users.

References

Agarwal, A. and A. Lavie (2008). METEOR, M-BLEU, and M-TER: Evaluation Metrics for High Correlation with Human Rankings of Machine Translation Output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, June 2008, pp. 115-118.

ASTM International (2016). *ASTM 2575 Standard Guide for Quality Assurance in Translation.* Retrieved February, 2016, from http://www.astm.org/Standards/F2575.htm.

Bojar, O., C. Federmann, B. Haddow, P. Koehn, M. Post, and L. Specia (2016). Ten Years of WMT Evaluation Campaigns: Lessons Learnt. In *Language Resources Evaluation Conference Proceedings*. Retrieved June, 2016, from http://lrec2016.lrec-conf.org/en/about/conference-proceedings/

Lommel, A., M. Popović, A. Burchardt (2014a). Assessing Inter-Annotator Agreement for Translation Error Annotation. Workshop on Automatic and Manual Metrics for Operational Translation Evaluation. *Proceedings of the Language Resources Evaluation Conference*, 2014. Retrieved June, 2016 from http://www.lrec-conf.org/proceedings/lrec2014/index.html.

Lommel, A., Uszkoreit, H., and Burchardt, A. (2014b). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista tradumàtica: technologies de la traducció*, 12, December 2014. ISSN 1578-17559.

Melby, A. (2016). A Spectrum from All MT at the Left to All HT at the Right End.Two-page Spec-Oriented Description for Spectrum v 4b.

O'Brien, S. (2012). Towards a Dynamic Quality Evaluation Model for Translation. *The Journal of Specialized Translation*, 17 January, pp. 55-77.

Rehm, G., A. Burchardt, O. Boja, C. Dugast, M. Federico, J. van Genabith, B. Haddow, J. Hajič, K. Harris, P. Koehn, M. Negri, M. Popel, L. Specia, M. Turchi, and H. Uszkoreit, (2016). Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem. In *Language Resources Evaluation Conference Proceedings*. Retrieved June, 2016, from http://lrec2016.lrec-conf.org/en/about/conference-proceedings/

Snow, T. (2015). *Establishing the Viability of the Multidimensional Quality Metric*. Dissertation, Brigham Young University. Paper 5593.

Translation Automation User Society (2016). The TAUS Dynamic Quality Dashboard: An Industry Collaborative Platform for Translation Quality and Tracking. Retrieved August, 2016 from http://www.slideshare.net/TAUS/quality-dashboard-an-industry-collaborative-platform-for-translation-quality-measurement-and-tracking-achim-ruopp-and-jaap-van-der-meer-taus/

U.S. Government Interagency Language Roundtable (2016). ILR Skill Level Descriptions for Translation Performance. Retrieved May, 2016 from http://www.govtilr.org/skills/AdoptedILRTranslationGuidelines.htm.

Uszkoreit, H. and A. Lommel. Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment. Retreived October, 2016 from http://www.qt21.eu/launchpad/sites/default/files/MQM.pdf

Van Ess-Dykema, C., J. Phillips, F. Reeder, L. Gerber (2011). Paralinguist Assessment Decision Factor for Machine Translation Output: A Case Study. Retrieved October, 2016 from http://www.mt-archive.info/AMTA-2010-VanEss-Dykema.pdf