Why is it So Hard to Compare Translation Evaluations and How Can Standards Help?

AMTA 2020

9 October 2020

Jennifer DeCamp

jdecamp@mitre.org

Why is it so Hard to Develop Comparable Translation Evaluations and How Can Standards Help?

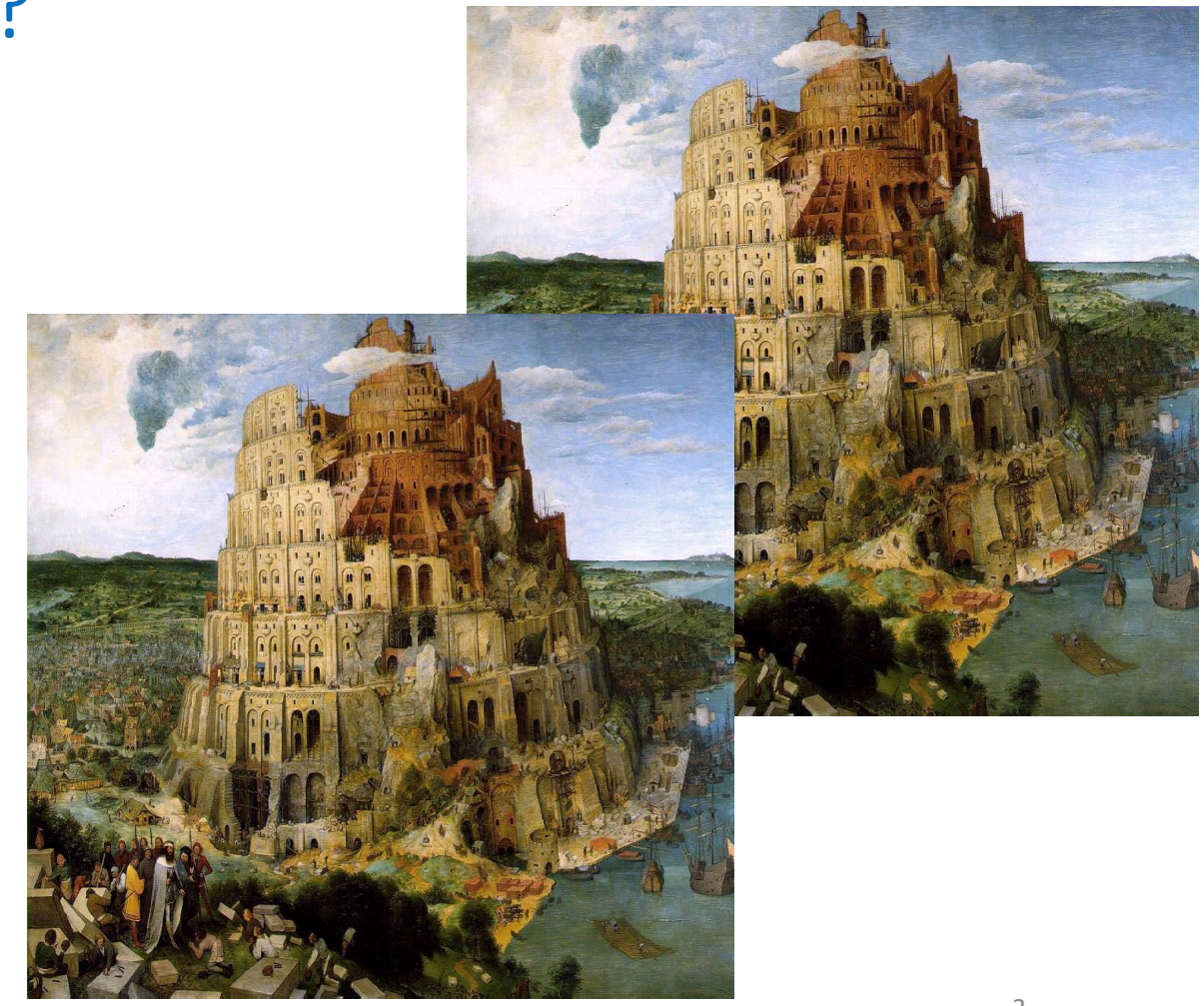
There are several standards and guidelines that may be relevant to MT and that are in use or anticipated to be ready for use this year, including *ASTM 2475 Translation Quality Requirements*, *ASTM WK46396 Analytic Evaluation of Translation Quality*, *ISO 17100 Translation services — Requirements for translation services*, and the Interagency Language Roundtable Skill Level Descriptions for Translation Performance.

This presentation reviews these standards and guidelines and discusses how they can be applied to evaluation of MT, whether HT, MT, or some combination. Such comparisons may include: a new version of MT with the previous version; one company's MT with that of another company, one product with another product; a language service provider's performance in one year vs. another, or one organization with another. The presentation also addresses gaps and provides recommendations, including to for become involved with improving these standards and thus improving MT evaluation.

Why Is It So Hard to Develop Comparable Translation Evaluations?

1. Variation in the translations
   - Languages, dialects, registers, domains
   - Genres
   - Cultural information
   - Processes
   - Tools
   - Purposes
   - Terminology
   - Need for reliability
   - Etc.

2. Variation in translation evaluations
   - Purposes
   - Requirements
   - Methods
   - Tools
   - Terminology
   - Need for reliability
   - Etc.



Purposes for Translation Evaluation

1. Development progress and direction
2. Acquisition
   - Write and administer contracts
     - *48 Code of Federal Regulations (CFR) § 15.101-2 - Lowest price technically acceptable source selection process*: The evaluation factors and significant subfactors that establish the requirements of acceptability shall be set forth in the solicitation.
   - Make decisions
     - Obtain
     - Upgrade
     - Replace
3. Management
   - Deploy resources
   - Determine performance
     - Track performance over time
     - Benchmark
4. Determine quality of deliverable
   - Send it for revision
   - Deliver it to the customer

Types of Evaluation

- Product
  - Reference
    - Human translation
  - No reference
    - Impact in workflow (e.g., impact to entity extraction)
    - Detailed error analysis
      - *ATA Certification Scoring*
      - *ASTM WK46396 Analytic Evaluation of Translation Quality*
- Process
  - *ISO 17100 Translation services — Requirements for translation services*
  - *ASTM 2575 Requirements for Translation Evaluation*
  - Skill descriptions by ILR, DLPT, ACTFL, etc.
- Outcome
  - Impact of the translation (e.g., in comparison with source text)

Methods and Tools

"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, tools and data sets.*
- *As a consequence, it is difficult to reproduce, interpret, and compare evaluation results"* *(G. Rehm et al, 2016)*

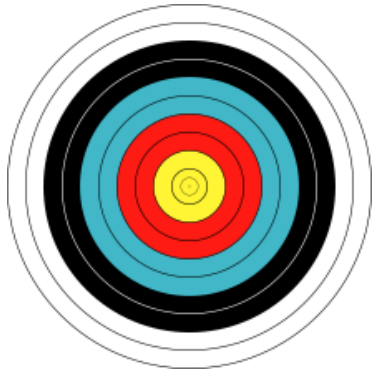
Standards

- *ASTM WK46396 Analytic Evaluation of Translation Quality*
- *ASTM 2575 Requirements for Translation Evaluation*
- ISO and ASTM efforts in terminology management

Meet Customer Requirements

Translator or Language Service Provider	“What the customer wants”
Customer	“100% accuracy, fast, cheap”
MT Developer	“As good as a human”

Quality/Target



- Announcements of NMT equaling or exceeding human performance
  - As good as human translators (i.e., humans hired from a translation company)
  - “If they did that work at my company, they wouldn’t be working for me for long” (owner of translation company at last AMTA meeting)
- But anyone can self-declare as a translator; any company can self-declare as a translation company
- And in the above methods, evaluators are not doing translator tasks
- Plus extensive issues with use of reference translations
  - Can’t always use a reference translation (e.g., to determine whether a translation is ready to submit to a client)
  - Get different results with different reference translations or different number of reference translations or references from a different part of the translation pair
  - Numbers often not meaningful at quality levels needed for deliverables
  - Usually reviewed by people with no understanding of context
  - Provides little or no information on WHAT is wrong
  - May penalize for different terminology or word order
  - Etc.

A. Lommel (2016)

What Humans?

- American Translators Association (ATA) Certification
  - Focused on perhaps too high a level
  - Too time-consuming for most applications
  - Not available for many languages
  - Directory of translators with certification and resumes listed on ATA home page http://www.atanet.org
- *ISO 17100:2015 Translation Services – Requirements for Translation Services*
  - Human translation (no technology), with an amendment to cover requirements in the U.S., Canada, and several other countries
  - CHF 118 (=UDS 128.40)
  - Britain's Institute of Translation and Interpreting (ITI) has created a translator "qualification" for meeting requirements for translators
  - U.S. and others have proposed having a new standard to better meet the needs for certification
- Defense Foreign Language Proficiency Test (DLPT) and American Council on the Teaching of Foreign Languages (ACTFL) Scores

Institute of Translation and Interpreting

Interagency Language Roundtable
Skill Descriptions for Translation Performance

- **Level 2+ (Limited Performance)**

  Can render straightforward texts dealing with everyday matters that include statements of fact as well as some judgments, opinion, or other elements which entail more than direct exposition, but do not contain figurative language, complicated concepts, complex sentence structures, or instances of syntactic or semantic skewing.

- **Level 3 (Professional Performance)**

  Can translate texts that contain not only facts but also abstract language, showing an emerging ability to capture their intended implications and many nuances. Such texts usually contain situations and events which are subject to value judgments of a personal or institutional kind, as in some newspaper editorials, propaganda tracts, and evaluations of projects.

Interagency Language Roundtable
Skill Descriptions for Translation Performance

Level 4+ (Professional Performance Plus)

- Can successfully apply a translation methodology to translate texts that contain highly original and special purpose language (such as that contained in religious sermons, literary prose, and poetry). At this level, a successful performance requires not only conveying content and register but also capturing to the greatest extent all nuances intended in the source document. Expression is virtually flawless.

Level 5 (Professional Performance)

- Can successfully translate virtually all texts, including those where lack of linguistic and cultural parallelism between the source language and the target language requires precise congruity judgments and the ability to apply a translation methodology. Expression is flawless.

How can Standards Help in MT Evaluation?

1. Can provide a better understanding of the information needed by decision makers
   - Target level
   - Requirements
2. Can help provide a broader framework for structuring an evaluation
   - Broader information
   - Authority
3. Can improve communication through this framework and through standardized terminology

Recommendations

1. Work towards a common framework
   - Employ standardized terminology
   - Employ interoperability standards for exchange of data
   - Carefully test and document results
2. Become more specific
   - Look at customer requirements for translation and translation evaluation
   - Reduce ambiguity re “human”
3. Educate the customer on
   - Requirements and opportunities training
   - On the impact of evaluation methods on certain types of translation
4. Participate in developing standards
   - ASTM
   - ISO

References

48 Code of Federal Regulations (CFR) § 15.101-2 - Lowest price technically acceptable source selection process. Retrieved September 25, 2020 from: https://www.law.cornell.edu/cfr/text/48/15.101-2

M. Dilllinger (2016). *MT Escaped from the Lab: Now What?* AMTA 2016. Retrieved September 25, 2020 from: https://amtaweb.org/wp-content/uploads/2016/10/Dillinger_AMTA2016Keynote_dist.pdf

G. Rehm, A. Burchardt, O. Bojar, C. Dugast, M. Federico, J. van Genabith, B. Haddow, J. Hajic, K. Harris, P. Koehn, M. Negri, M. Popel, L. Specia, M. Turchi, and H. Uszkoreit (2016). *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem,* Language Resources Evaluation Conference (LREC) Workshop, 24 May 2016. Retrieved September 25, 2020 from: http://www.cracking-the-language-barrier.eu/mt-eval-workshop-2016/

A. Lommell (2016). *Blues for BLEU: Reconsidering the Validity of Reference-Based MT Evaluation*, LREC Workshop on *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, 24 May 2016. Retrieved September 25, 2020 from http://www.cracking-the-language-barrier.eu/mt-eval-workshop-2016/.

American Translators Association (2020). *Directory*. http://www.atanet.org

ASTM (2020). *ASTM WK46396 Analytic Evaluation of Translation Quality.* Draft Standard. Copies available from ASTM for review.

ISO (2015). *ISO 17100 Translation services — Requirements for translation services. ISO Standard.* Retrieved September 25, 2020 from: https://www.iso.org/standard/59149.html

ASTM (2020). *ASTM 2575 Requirements for Translation Evaluation: Standard Practice.* Draft standard anticipated for publication 2020. Copies available from ASTM for review.

ASTM (2014). *ASTM 2575 Requirements for Translation Evaluation.. Standard Guide.* Retrieved September 25, 2020 from: https://www.astm.org/Standards/F2575.htm

Interagency Language Roundtable (ILR), ILR Skill Descriptions for Translation Performance. Retrieved September 25, 2020 from https://www.govtilr.org/Skills/AdoptedILRTranslationGuidelines.htm

Images