Engaging Content
Engaging People

Machine Translation Evaluation: know your essentials

Dr. Sheila Castilho

Tutorial at the RANLP'19

September 1st, 2019

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund
SFI Science Foundation Ireland

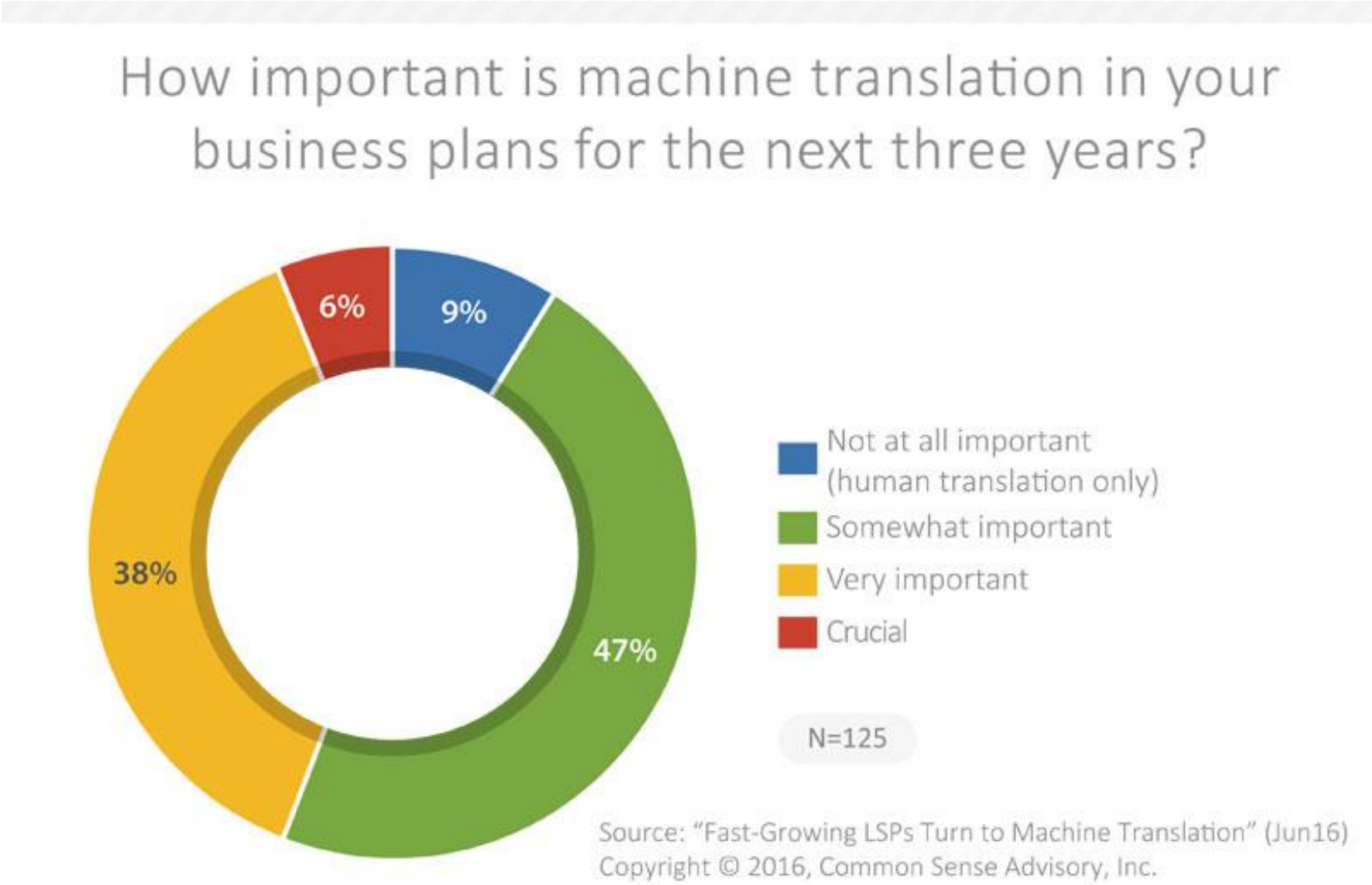
The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Outline

1. Machine Translation Market: now and then
2. Why MTEval?
3. Types of MTEval
    a) AEMs
        i. QE
    b) HEMs
        i. Inter-Annotator Agreement
        ii. User Evaluation
        iii. Test Suites
4. Further Developments

MT is important

- 50 of the top 100 language service providers (Common Sense Advisory 2016)



Why do we need evaluation?

- Evaluation provide data on whether a system works and why, which parts of it are effective and which need improvement.

- Evaluation needs to be **honest** and **replicable**, and its methods should be as rigorous as possible.

Why do we need evaluation? A bit of history...

- ALPAC Report (1964)
- Generated a long and drastic cut in funding (especially in MT)
- Evaluation was a forbidden topic in the NLP community (Paroubek et al 2007)

First question: what is quality?

➢ Evaluation is a complex problem

- What does quality mean?
  - Fluent? Adequate? Both? Easy to post-edit? Usable? All of them? None of them?

Quality for whom/what?

- Why are you evaluating the MT system?
  - End-user (gisting vs dissemination)
  - Post-editor (light vs full post-editing)
  - Other applications (e.g. Cross Lingual IR)
  - MT-system (tuning or diagnosis for improvement)

Quality for whom/what?

- Why are you evaluating the MT system?
  - End-user (gisting vs dissemination)
  - Post-editor (light vs heavy post-editing)
  - Other applications (e.g. CLIR)
  - MT-system (tuning or diagnosis for improvement)

How to evaluate MT?

- A few methods
- Automatic evaluation
  - Automatic evaluation metric (AEMs)
  - Automatic classifications

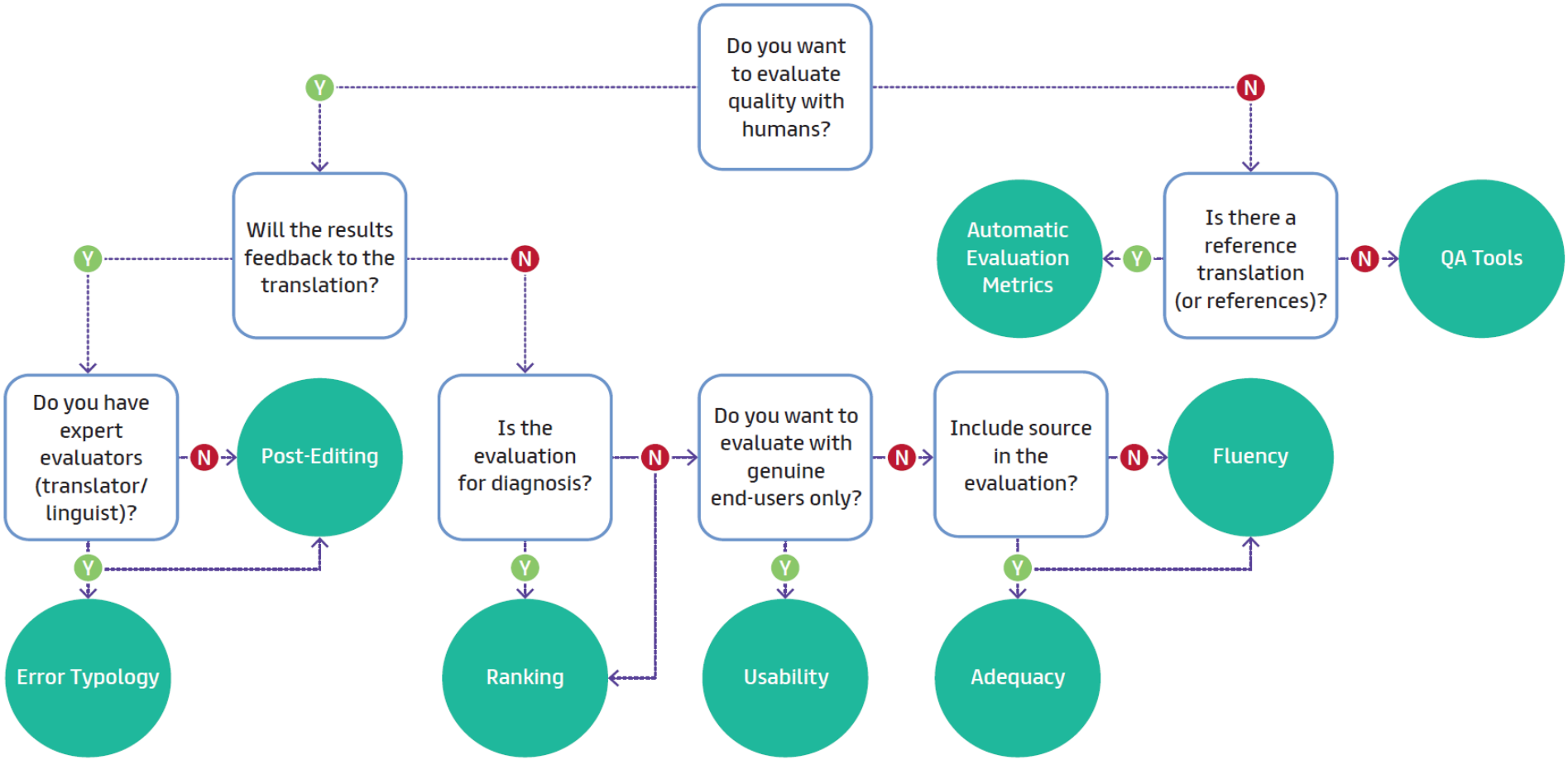
- Human Evaluation
  - Human evaluation metrics (HEMs)
  - Professional translators/bilinguals/crowd
  - User Evaluation
    - Usability, reading comprehension (UEMs)

How to evaluate MT?

- Quality Assessment tools (QA)
  - (Semi)automatic
  - Heavily (still) applied in industry

- Quality Estimation (QE)
  - “Not really evaluation”
  - reference translation is not available

Translation Evaluation Flowchart



Moorkens, J., Castilho, S., Gaspari, F., Doherty, S (Ed.). (2018) *Translation Quality Assessment: From Principles to Practice.* Heidelberg: Springer.

Automatic Evaluation Metrics (AEMs)

➢ Interdisciplinary

- WER (speech recognition – MT)
- ROUGE (text summarization – MT)
- F-Measure (IR – many other areas)

Automatic Evaluation Metrics (AEMs)

What is an AEM?

- a computer program:
  - input: translation output and reference translation(s)
  - output: a numerical score related to their similarity

Automatic Evaluation Metrics (AEMs)

What is an AEM?

- a computer program:
  - input: translation output and reference translation(s)
  - output: a numerical score related to their similarity
  - Evaluation without references => Quality Estimation
- usual methods for comparison
  - n-gram matching
    - F-scores, BLEU, METEOR
  - edit (Levenshtein) distance
    - WER, (H)TER

AEMs based on n-gram matching

- BLEU
  - geometric mean of 1-, 2-, 3- and 4-grams
  - precision + brevity penalty instead of recall (for penalising too short segments)

- METEOR
  - flexible unigram recall
  - does not penalise (too hard) common stems, synonyms and paraphrases

AEMs based on Character level

take into account characters instead/in addition to words

ref I t w i l l b e a s o r t o f b r i d g e .

mt I t w i l l s o r t o f b r i d g e b e .

Metrics

- chrF: character n-gram F-score possibly extended by word 1-grams and 2-grams
- characTER: character based extended edit distance TER unmatched words are compared on character level
- BEER: combination of word n-grams, character n-grams, word order permutations

AEMs based on Edit (Levenshtein) distance

Minimum number of edits to transform translation output to the reference translation

edit types:

- substitution: replace one word with another
- deletion: a word is missing, it should be added
- insertion: a word is inserted, it should be removed

Metrics

- Word Error Rate (WER) – normalised Levenshtein distance
- Translation Edit Rate (TER) (Snover et al, 2006) – WER extended by block shift (reordering) cost

AEMs based on Edit distance - TER

- TER – Translation Edit Rate – and HTER – Human-targeted edit rate

- TER uses the MT as the hypothesis and the HT as reference
  - The higher the score, the higher number of edits for the MT to come closer to the HT

- HTER can have hypothesis and reference **interchanged** (Do Carmo 2019):
  - The higher the score, the more edits performed by the translator
  - MT hypothesis and PE reference: Shows the error in the MT
  - PE hypothesis and MT reference: Shows the edits performed by the translator

AEMs based on Edit distance - HTER

HTER can have hypothesis and reference **interchanged**:

UNEDITED (MT)	(POST)EDITED
This sentence has a redundant **superfluous** word.	This sentence has a redundant word.
In this sentence, a is missing.	In this sentence, a **word** is missing.
This sentence has an **incorrect** word.	This sentence has a **corrected** word.
In this sentence, all are correct **words**, but one is in the wrong position.	In this sentence, all **words** are correct, but one is in the wrong position.

- MT hypothesis and PE reference (error in MT):
  - **Insertion, Deletion**, Substitution, Shift
- PE hypothesis and MT reference (edits in PE):
  - **Deletion, Insertion**, Substitution, Shift

Other AEMs

variations and combinations (word and character level, n-grams and edit distances, evaluating on POS tags, morphemes, discourse markers, etc.)

Keep an eye for the WMT Shared task metrics![1]

[1] http://www.statmt.org/wmt19/metrics-task.html

What about Quality Estimation (QE)?

- Not a measure of quality: no comparison against a reference
- Provides an **estimate** on the quality of translations on the fly

Quality is data driven:

- Can the translation be published as it is?
- Can a reader get the gist?
- Is it worth post-editing it?
- How much effort to fix it?

(Specia, 2016)

What about Quality Estimation (QE)?

- Features extracted from examples of translation and source
  - Source -> complexity features (i.e. how hard it is to translate?): sentence length, common words (frequency of words)
  - Translation-> fluency features: grammatical (i.e. grammar checker), sequence of words
  - Source+ translation -> adequacy features (i.e. difference in length)
  - PEs and human annotated data can also be used

What about Quality Estimation (QE)?

- Features can be tailored and extracted depending on the definition of quality
  - How much effort to fix the translation?
    - PE time
  - Can the translation be published as it is?
    - Adequacy, fluency scores
  - Can a reader get the gist?
    - General adequacy, fluency from sample translations

Automatic Metrics: Pros and Cons

- Time efficient
- Inexpensive
- Consistent

But…

Exactly matches are not a good way to evaluate:

- many ways to translate the meaning of the source text
- AEMs would need many different reference translations for the same source text
- AEMs don't tell us *what* is happening
  - do not give any details about actual translation errors

Automatic Metrics: Pros and Cons

What overall evaluation scores cannot answer?

- What is a particular strength/weakness of the system?
- What does a certain modification of a system improve exactly?
- Does a worse-ranked system outperform a better-ranked one in any aspect?

Deeper analysis is needed!

Human Translation Quality Assessment (TQA)

- Why evaluate machine translation with humans?
  - More detailed evaluation
  - Assess complex linguistic phenomena
  - Deeper analysis of system's performance
  - Feedback to the MT system
  - Diagnosis

Human Evaluation is essential because it avoids things like…

The hype

Technology

"Nearly Indistinguishable From Human Translation"—Google Claims Breakthrough

by Florian Faes on September 27, 2016

The Google team working on neural

Microsoft Research's AI system achieves human parity in translating news from Chinese to English

by Pradeep @pradeepviswav Mar 14, 2018 at 14:13 GMT

theDigitalExperience ▸ Insights Worth Sharing

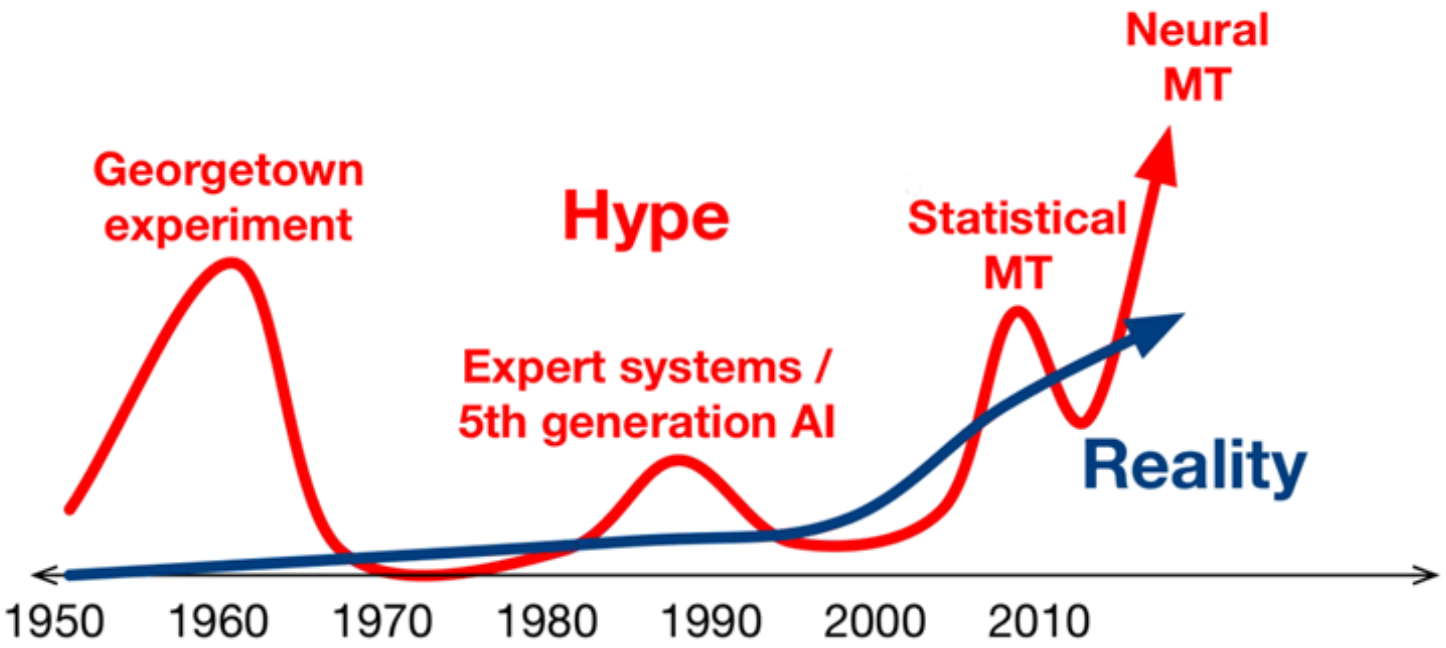
Blogs Home | Customer Experience | Content Management | Language

SDL Cracks Russian Neural Machine Translation

By Kirti Vashee / August 3, 2018

The SDL Machine Translation (MT) research team recently announced that our latest machine learning innovations and development strategies with Neural MT have resulted in a breakthrough that clearly demonstrates a significant and substantial leap forward. When testing our Russian to English MT system, the output, when measured against an extensive suite of comparative experiments to verify and validate the outstanding

Hype and Reality



(Philipp Koehn, Omniscien Webinar 2017)

Human Translation Quality Assessment (TQA)

- most common metrics (Castilho et al 2018):
  - Adequacy and fluency,
  - Error classification
  - Post-editing
  - Ranking
- secondary measures are: readability, comprehensibility, usability, acceptability of source and target texts.
- carried out by professional and amateur evaluators.
- **performance-based** measures and **user-centred approaches** are more recent additions.

Ranking

- Compares two or more translations
  - Comparison can be between translation from different systems or human translation
- Draws may or not be allowed
  - May also be divided into “equally good” or “equally bad”

Ranking

- evaluation:

"which of these translations are better?" or " choose your preferred translation"

001/2001

Given the translations by more than two MT systems, the task is to rank them: - Rank a system A higher (rank1) than B (rank2), if the output of the first is better than the output of the second. - Rank both systems equally, A rank1 and B rank1, if the outputs are of the same quality - Use the highest rank possible, e.g. if you've three systems A, B and C, and the quality of A and B is equivalent and both are better than C, then do: A=rank1, B=rank1, C=rank2. Do NOT use lower rankings, e.g.: A=rank2, B=rank2, C=rank4.

28岁厨师被发现死于旧金山一家商场 近日刚搬至旧金山的一位28岁厨师本周被发现死于当地一家商场的楼梯间。
— Source

28-Year-Old Chef Found Dead at San Francisco Mall
A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week.
— Reference

○ Rank 1 ○ Rank 2 ○ Rank 3
The 28-year-old chef was found dead at a San Francisco mall
— Translation 1

○ Rank 1 ○ Rank 2 ○ Rank 3
28-year-old chef found dead in San Francisco mall
— Translation 2

○ Rank 1 ○ Rank 2 ○ Rank 3
28-year-old chef found dead in a shopping mall in San Francisco
— Translation 3

Submit Reset Flag Error

Ranking

Why is ranking useful for MTEval?

- It tells us whether the assessed system is improving compared to the baseline (diagnosis evaluation)
- It tells us which system is more suitable for a specific project

Adequacy

- also known as “accuracy” or “fidelity”
- Focus on the **source** text
- “the extent to which the translation transfers the meaning of the source text translation unit into the target”

- Likert scale:

1. None of it
2. Little of it
3. Most of it
4. All of it

Adequacy

- Why is Adequacy useful for MT evaluation?
  - It tells us how much of the source message has been transferred to the translation
  - Sometimes you are only interest in the meaning of the source sentence

Fluency

- also known as intelligibility
- focuses on the target text
- “the flow and naturalness of the target text unit in the context of the target audience and its linguistic and sociocultural norms in the given context”
- Likert scale:

1. No fluency
2. Little fluency
3. Near native
4. Native

Fluency

- Why is Fluency useful for MT evaluation?
  - It tells if the message is fluent/intelligible (i.e. sounds natural to a native speaker) or if it is “broken language”.

Adequacy-Fluency

- Adequacy and Fluency generally go together
  - But sometimes you may want to prioritise one over the other
    - Technical documentation may require more adequacy

Pos-editing (PE)

- The “term used for the correction of machine translation output by human linguists/editors” (Veale and Way 1997)

- “checking, proof-reading and revising translations carried out by any kind of translating automaton”. (Gouadec 2007)

- Common use of MT in production – over 80% of Language Service Providers now offer post-edited MT (Common Sense Advisory 2016)

Light Post-editing vs Full Post-editing

Light Post-Editing	Full Post-Editing
Essential corrections only	More corrections leading to higher quality
Quick turn-around	Slower Turn-around
General texts that are needed urgently	Aim at general audience (dissemination, outbound)
Internal and perishable use	Texts that corresponds to human quality
Correct blatant errors without considering style	Not only blatant errors, but all errors and style
Types: emails, reports, meeting agendas, technical reports, user forums, chat-rooms	Types: material to be published, software, technical documentation

Light Post-editing vs Full Post-editing

- Is this distinction really useful for a translator?

  "Light PE does not exist"

- Guidelines are essential

  - TAUS PE guidelines

Measurement of PE effort

From Krings' book Repairing Texts (2001)

- Temporal effort
  - Throughput, the amount of time spent post-editing
  - Often expressed in words/second

- For MT Eval – faster better means better MT output?
  - productivity

Measurement of PE effort

- Technical effort
  - The number of edit operations made
  - Often approximated using hTER automatic metric
  - For MTEval – fewer edits mean better MT
    - Correlates with time effort = productivity
  - HTER
    - PE as reference
    - PE as hypothesis

Measurement of PE effort

- Cognitive effort
  - May be measured in several ways
  - In DCU we often use eye-tracking

  - For MT Eval – less cognitive effort means better MT output
  - Cognitive effort has been correlated to other HEMs

- Why use post-editing for Machine Translation evaluation?

  - Assess usefulness of MT system in production
  - Identify common errors
  - Create new training or test data

  - However, measurements of post-editing effort tend to differ between novice (students) and professionals, and crowd and professionals

Tools for PE

PET (Aziz, Castilho and Specia 2012)

CASMACAT (Alabau et al. 2013)

Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation

MATECAT (Federico et al. 2014)

Appraise (Federman, 2012)

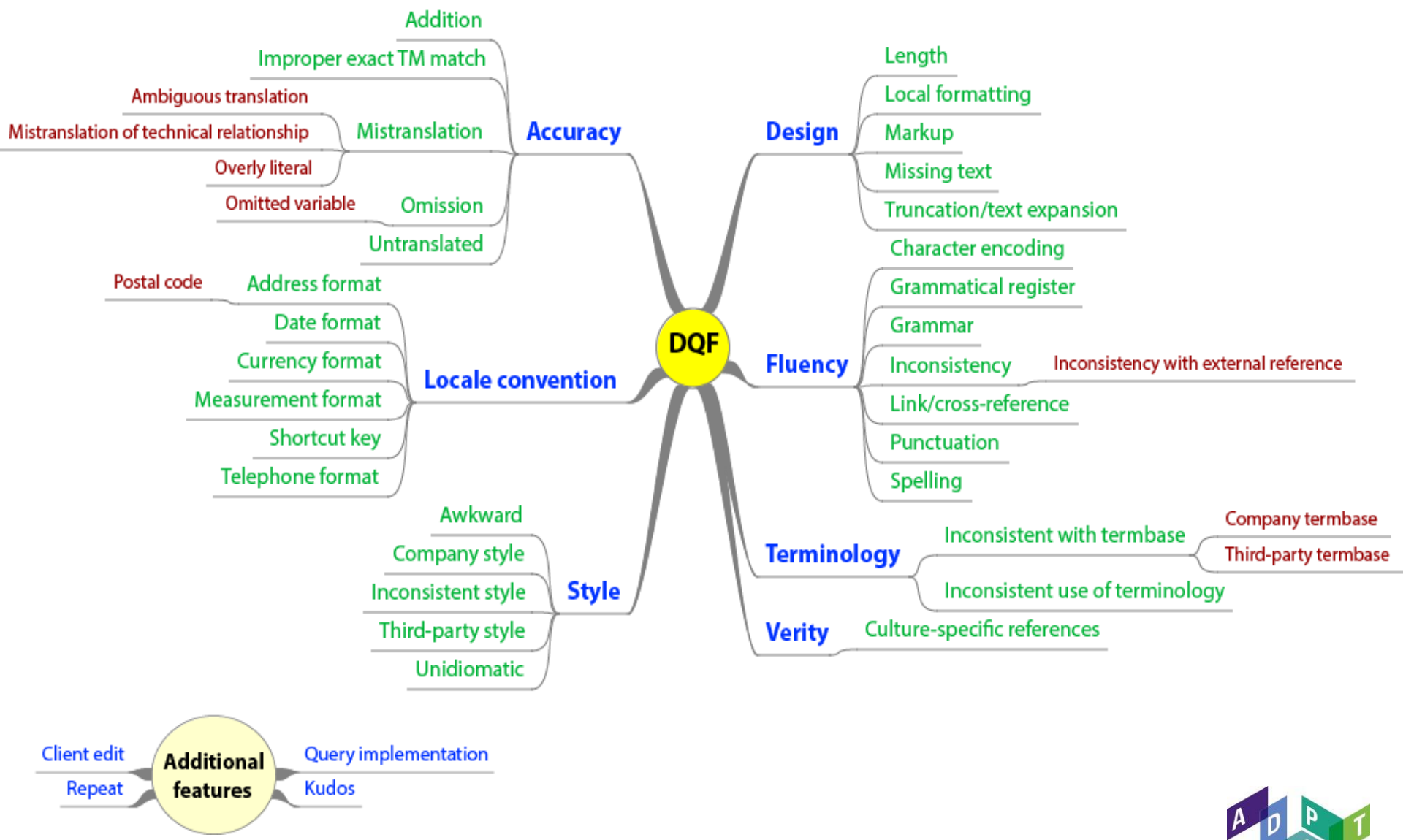
Error Classification

- Identify and classify errors in a translated text
- A few taxonomies have been proposed

- Vilar et al. (2006)
- Llitjós et al. (2005).
- Federico et al. (2014)
- Costa et al. (2015)
    - DQF – TAUS
    - MQM – QT21[2]
        - Harmonized

[2] http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

DQF / Multidimensional Quality Metrics



Error Classification

- Why use error taxonomies for translation evaluation?
  - Identify types of errors in MT or human translation
  - Detailed error report is useful for adjusting MT systems, reporting back to clients
  - LSPs use taxonomies and severity ratings to monitor translators' work

Error Classification

- More possible analyses:
  - relations between particular error types and user/post-editor preferences
  - the impact of different error types on different aspects of post-editing effort

- *However, error annotation is expensive*

- Automatic Error Classification has been Proposed (See Popovic 2018)

DQF / MQM Example

ST: Quando você faz avaliação humana dos sistemas, é mais provável que os seus resultados tenham mais peso.

MT: When you **make** **human systems evaluation**, it is more likely that **the** your results will have much more **weight.**

HT: When you do human evaluation **of the** systems, it is more likely that your results will have more credibility.

Errors:

- **Word order**
- **Mistranslation**
- **Literal Translation**
- **Extraneous function word**
- **Addition**
- **Missing**

Inter-annotator agreement (IAA)

Necessary because human annotation/evaluation is:

- Subjective
- Prone to errors (fatigue)
- Biased (preference for a label)
- Based on human-written guidelines (misinterpreted)

Inter-annotator agreement (IAA)

Results can:

- Identify improvements needed in annotation scheme
- Indicate usefulness of data
- Indicate replicability of data (e.g. clinical diagnoses)

Most used coefficient:

- Cohen's Kappa (weighted and non-weighted)
- Fleiss' Kappa

The ultimate goal is to identify and solve disagreements to have a more homogenous annotation

Usability

- Concept borrowed for human-computer interaction
- Real world problems
- Understand how end users engage with machine-translated texts or how usable such texts are.
  - Applied for different areas (video/text summarisation, UI, information retrieval, etc.).

Usability

- Why is Usability useful for MT evaluation?
  - identify what impact the translation might have on the final readers of the translation, including their satisfaction with the translation and products.
  - The users of the translation should be the ones who tell us if the final translation is acceptable

Challenge Test Set – Test Suites

"A challenge test set is a representative set of isolated or in-context sentences, each hand or (semi)automatically designed to evaluate a system's capacity to translate a specific linguistic phenomenon"

Generally:

it must concentrate on specific phenomena

it should represent well these phenomena

it should be of reasonable size

it should enable a straightforward evaluation

the phenomena are usually linguistically motivated

- Popovic and Castilho 2019. CTS MTSummit tutorial (see https://sites.google.com/view/challenge-test-sets-tutorial/home)

Challenge Test Set – Test Suites

- since 2015: revived in order to obtain more fine-grained qualitative observations about MT systems
- since 2017: expanded with the emergence of neural systems
- since 2018: "Additional Test Suites in News Translation Task“ at WMT

Challenge Test Set – Test Suites

An additional test - **and** it would not be difficult to prepare – would make the results stronger.

And now something completely different!

I'm a great actor, **and** you're a cheap producer.

Cathy thought she was going to win, **and** you pushed her.

Chris planned this trick **and** you carried it through.

Come to us not as a guest, **but** as a brother.

Convicted not of arson, **but** of some minor transgression.

Crime is not the reason **but** the consequence.

Don't talk, **but** do it now.

- Creation and Evaluation
  - (semi) automatic, manual

Further Developments

Human parity? What is it?

Hassan et al. (2018)

- **Definition 1**. If a **bilingual** human judges the quality of a candidate translation produced by a human to be equivalent to one produced by a machine, then the machine has achieved **human parity**.

Further Developments

- **Definition 2**. If there is no statistically significant difference between human quality scores for a test set of candidate translations from a machine translation system and the scores for the corresponding human translations then the machine has achieved human parity

Human Parity? Let's look at it a little bit closer

1. **Toral, Castilho, Hu and Way (2018):**

    a) Professionals vs bilinguals vs crowd

Why an issue:

non-expert evaluators lack knowledge of translation and might not be able to notice subtle differences that make one translation better than another (Castilho et al., 2017)

.

Human Parity? Let's look at it a little bit closer

b) Sentences were evaluated in isolation

Why an issue:

There are referential relations that go beyond the sentence level (Voigt and Jurafsky, 2012). These are disregarded in the evaluation.

Favouring MT?

Their MT system does not take into account intersentential context while human translators do.

Human Parity Disproved

Toral, Castilho, Hu and Way (2018):

- **Translationese:** if removed, evidence that human parity has not been achieved
- **Professional translators:** wider gap between HT and MT and higher IAA
- **Quality of human references:** issues seem to indicate that they were produced by non-expert translators and possibly post-edited

Laubli, Sennrich and Volk (2018):

- **Context**: stronger preference for human over MT when evaluating documents (vs isolated sentences)

Document-level Evaluation – next step!

WMT 2019 –

Direct Assessments (accuracy) with bilinguals for document level

Loads we don't know:

- how much context span is a "document level"
- effort to rate a whole document
- how to annotate errors in a doc-level evaluation

MTEval is essential

- AEM
  - Time efficient
  - Inexpensive

But no deep linguistic phenomena can be assessed (yet?), needs multiple human references to be less unfair

- HEMs

Can be expensive and time consuming

  - Assess linguistic phenomena
  - Feedback to the MT systems
  - Tell us what end users can/cannot do with the translation

Human Evaluation is not going anywhere

- Human evaluation avoids awkward situations…
  - Hype
- And back up good results!


tack
multţumesc
dhanyavād
takk
teşekkür ederim
mahalo
감사합니다
Баярлалаа
hvala
gracias
thank you
arigatô
dankie
Дякую
dziękuję
ευχαριστώ
dioch
dank
tänan
спасибо
merci
köszönöm
Obrigada
谢谢
kasih
תודה
choukrane
shukran
a dank
dĕkuji
terima
kiitos
Благодарам
grazie
gràcies
Շնորհակալություն
ačiū
Xièxiè
danke
ありがとう
благодаря
tak
Mh'gōi
ngiyabonga
paldies
grazzi
go raibh maith agat
kop khun
faleminderit
Ďakujem
хвала
nandri

References

Aziz, W., Castilho, S., and Specia, L. (2012). PET: a Tool for Post-editing and Assessing Machine Translation. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.

Alabau, V., Buck, C., Carl, M., Casacuberta, F., Garcıa-Martınez, M., Germann, U., et. al. (2014). Casmacat: A computer-assisted translation workbench. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. 25-28.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. Approaches to human and machine Translation Quality Assessment. In Translation Quality Assessment: From Principles to Practice, volume 1 of Machine Translation: Technologies and Applications, pages 9–38. Springer International Publishing, July 2018.

Costa A, Ling W, Luís T, Correia R, Coheur L (2015) A linguistically motivated taxonomy for machine translation error analysis. Mach Transl 29(2):127–161

Christian Federmann Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, Valletta, Malta, LREC, 5/2010

Federico M, Negri M, Bentivogli L, Turchi M (2014) Assessing the impact of translation errors on machine translation quality with mixed-effects models. In: Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, pp 1643–1653

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., ... & Massidda, A. (2014). The matecat tool. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations (pp. 129-132).

Krings, H. P. (2001). Repairing texts: empirical investigations of machine translation postediting processes. Kent State University Press, Kent, Ohio.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. https:// arxiv.org/abs/1803.05567.

Lommel, A. R. and DePalma, D. A. (2016). Europe's Leading Role in Machine Translation: How Europe Is Driving the Shift to MT. Technical report, Common Sense Advisory, Boston, USA.

Llitjós AF, Carbonell JG, Lavie A (2005) A framework for interactive and automatic refinement of transfer-based machine translation. In: Proceedings of the 10th conference of European Association for Machine Translation (EAMT2005), Budapest, pp 87–96

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of EMNLP*, pages 4791–4796, Brussels, Belgium.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers, pages 113– 123, Belgium, Brussels, October. Association for Computational Linguistics.

Maja Popovic. 2018. Error Classification and Analysis for Machine Translation Quality Assessment. In Translation Quality Assessment: From Principles to Practice. Morkeens, J., Castilho, S., Gaspari, F. and Doherty, S. (eds). pages 129–158. Springer International Publishing

Patrick Paroubek, Stephane Chaudiron, Lynette Hirschman. Principles of Evaluation in Natural Language Processing. Traitement Automatique des Langues, ATALA, 2007, 48 (1), pp.7-31.

Specia, L. Automatic Evaluation of Machine Translation: Moving Away from Word Matching Metrics. Gala Webinar, 2016.

Vilar D, Xu J, D'Haro LF, Ney H (2006) Error analysis of statistical machine translation output. In: Proceedings of 5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, pp 697–702