

In search of an acceptability/ unacceptability threshold in machine translation post-editing automated metrics

Lucía Guerrero
Machine Translation Specialist, CPSL

AMTA, October 2020

Why MT?

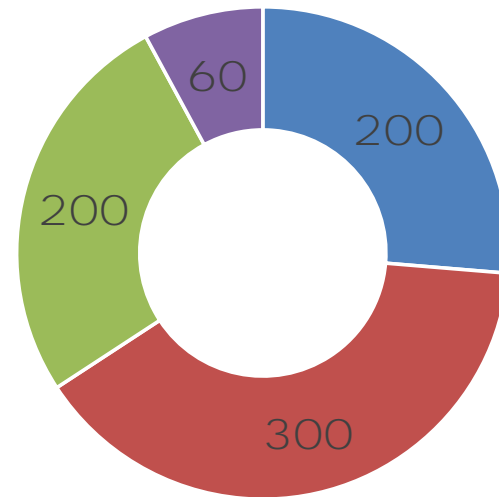


'Machines translate more in a day than all human translators on the planet combined can do in a year'

Nimdzi Research/TAUS, 2018



billion words/day



- Google Translate
- Alibaba
- Amazon
- eBay

MT main use cases and drivers



Translation for understanding:
raw MT / light postediting

E-commerce platforms
Forums and user reviews
Support pages
Communication apps



To cut costs and/or
improve deadlines:
light / full post-editing

MT at CPSL

SMT: Moses, ModernMT

NMT: Marian, 3rd-party platforms

RBMT: Apertium

Generic systems
and

Domain-based systems:

- Life sciences
- Medical devices
- Automotive
- Technical



Apertium



MODERN MT

MT evaluation



Holistic (adequacy/fluency) scoring
Perceived PE effort scoring



Reference-based metrics
(BLEU, edit distance, (H)TER...)



Productivity tests: post-editing time

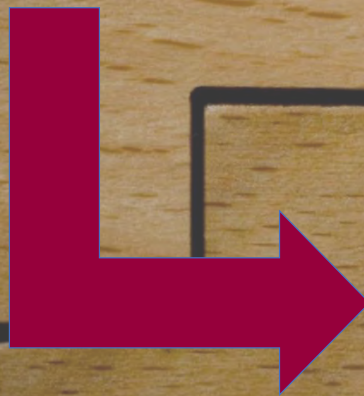


Analytical: all/main errors, categorized

Why...

... combining different types of evaluation?

- Human judgement alone is valuable but subjective
- Metrics alone are not enough




Combined metrics give
meaningful information

Why...

... searching for an acceptability threshold?

- Define goals when training systems
- Know when to retrain a system
- Cherry-picking projects for MT
- Avoid discussions on remuneration



What % of edit distance is acceptable/unacceptable for post-editing?

Previous studies

On acceptability:

- Castilho, S. (2016): “Measuring Acceptability of Machine Translated Enterprise Content”. Dublin City University, Dublin, Ireland.

On correlation between automated metrics and human judgement:

- Fomicheva, M.; Specia, L. (2019); “Taking MT Evaluation Metrics to Extremes: Beyond Correlation with Human Judgments”. On *Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, USA.
- Scarton, C.; Forcada, M.; Esplà-Gomis, M.; Specia, L. (2019): “Estimating post-editing effort: a study on human judgements, task-based and reference-based metrics of MT quality”. Proceedings of IWSLT 2019, Hong Kong, China.

Hypothesis:

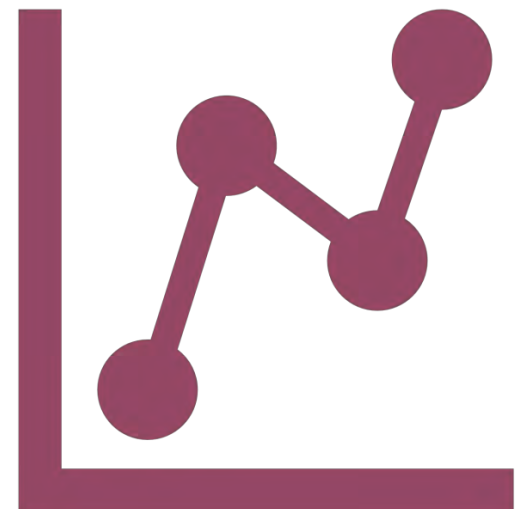


50% is too high as an
edit distance threshold
to define acceptability
of MT raw output

Description of study



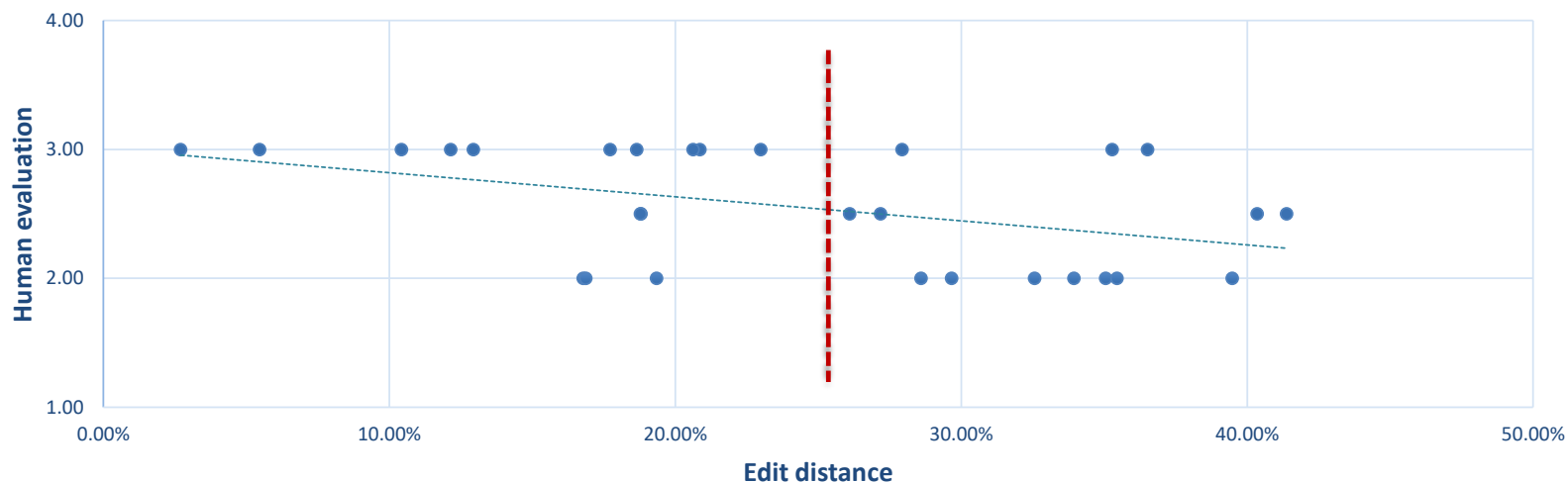
- ❑ 29 evaluations
 - ❑ Automated metrics: edit distance (Levenshtein algorithm from nltk.metrics)
 - ❑ Human evaluation after post-editing: PE effort perceived (1-4 Likert scale)
- ❑ 3 MT systems: Marian, Google Translate Basic and GT Advanced
- ❑ Evaluators' profile: professional post-editors
- ❑ 10 language combinations and 6 subject areas
- ❑ Limitations:
 - ❑ Usually only 1 post-editor (and evaluator) per project
 - ❑ Likert scores are subjective
 - ❑ Metrics result from comparing with the final version (sometimes there is an extra review)
 - ❑ Too few evaluations



Correlation table



Distribution between human scores and edit distance



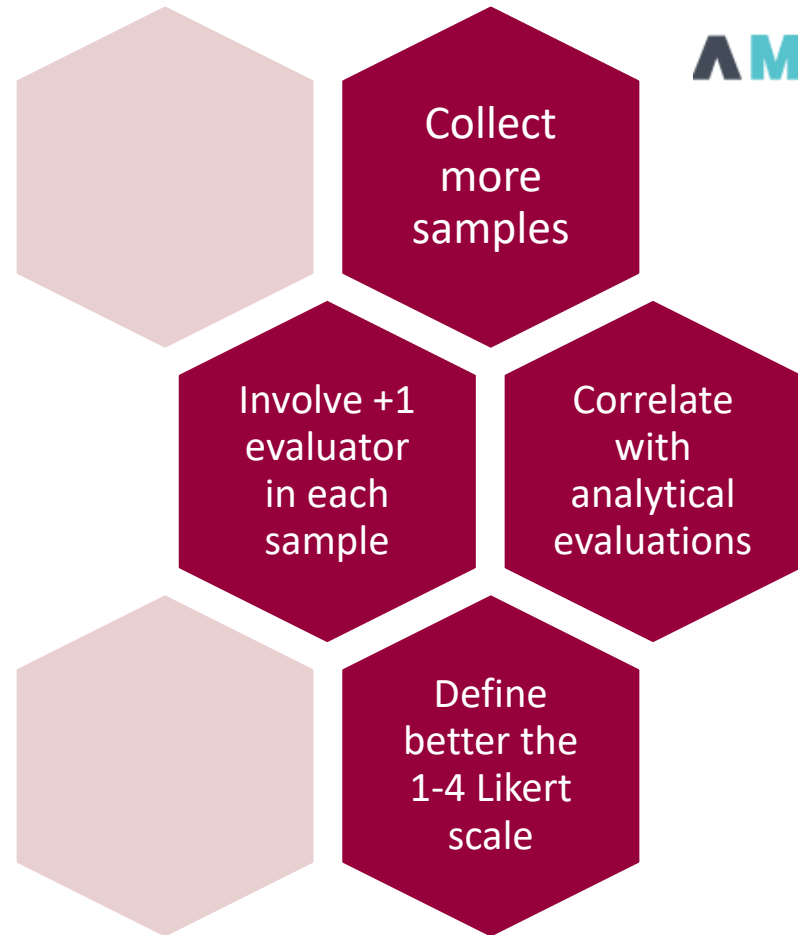
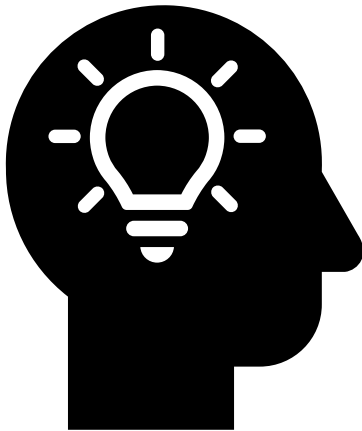
Interpretation

- Raw MT output scores: 2-3
- Most edit distances: 15%-45%
- Correlation? A high edit distance usually has a low score, and the other way around (but note the exceptions)
- According to the specific comments, 3 is usually related to good quality, whereas 2 seems to be closer to unacceptability



Possible interpretation:
with an edit distance
> 30%, post-editors
expect an improvement
of the raw MT output
in the next job

Ideas for further study



Questions?

Thank you!



CPSL Barcelona
Tel +34 93 445 17 63
info-spain@cpsl.com

CPSL Madrid
Tel +34 91 787 48 61
info-spain@cpsl.com

CPSL Germany
Tel +49 (0)71 41 - 97 00 006
info-germany@cpsl.com

CPSL UK
Tel (+44) 207 993 4550
info-uk@cpsl.com

CPSL USA
Tel 1 (617)-399-8194
info-usa@cpsl.com

cpsl.com