Assessment of Machine Translations of Survey Questions and Response Scales

Using metrics to evaluate Machine Translation Quality

European Social Survey
upf. Universitat Pompeu Fabra Barcelona
RECSM Research and Expertise Centre for Survey Methodology

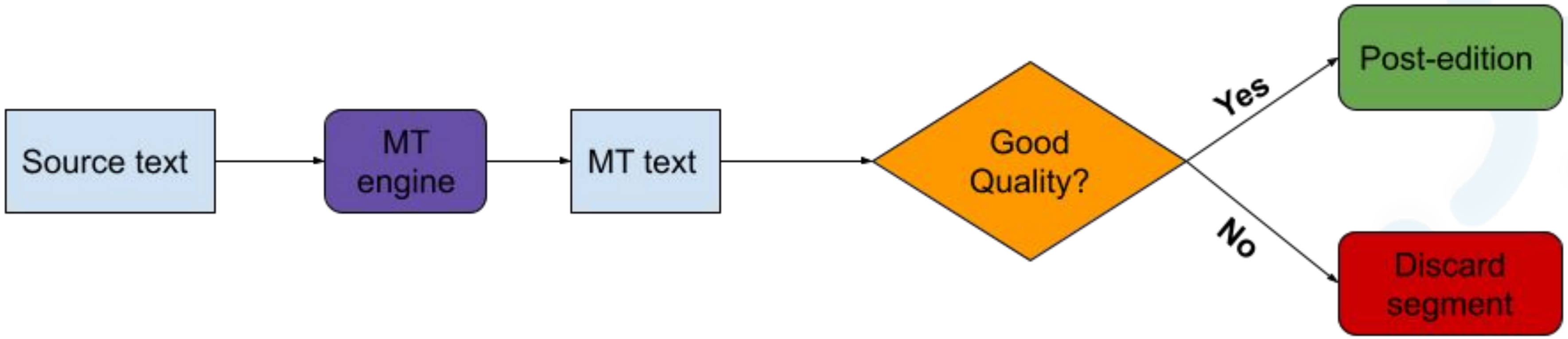
Danielly Sorato, UPF
Diana Zavala-Rojas, UPF
Veronika Keck, GESIS
Dorothée Behr, GESIS
Brita Dorer, GESIS

European Survey Research Association (ESRA)
July 09, 2021

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

Main objective



- **Machine Translation (MT) evaluation is an important step that should be added <u>prior to Post-editing</u>**
  - **If a given MT output has bad quality, it may be more troublesome to fix it rather than start a new translation from scratch**

Evaluate the quality of MT outputs in this experiment from a computational perspective

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Specific objectives

1. To investigate the quality of machine translated sentences in against the review version of the baseline treatment (fully human pipeline)
   a. Using sentence similarity metrics
   b. Using MT evaluation metrics

However, choosing a reference translation can be problematic
- Scores biased to the vocabulary and phrasing of the reference
- There are cases where a reference translation is not available (e.g. new survey items)

Therefore the MT evaluation paradigm is changing to…

2. Evaluating the quality of the machine translated sentences using a Quality Estimation (QE) model
   - No need for reference translation
   - Models trained on MT outputs and their post-editions

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

1a. Similarity metrics

- **Levenshtein distance (lexical): the minimum path of necessary edits to transform string (words, sentences) into another.**

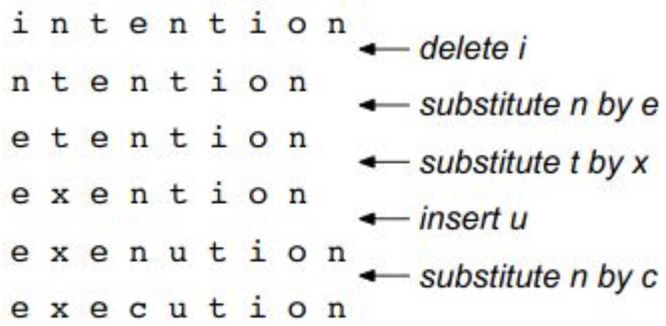


Image from Speech and Language Processing (3rd ed. draft) https://web.stanford.edu/~jurafsky/slp3/

- **Fuzzy word match taking order into account (syntactic). Percentage of words that are a matched in the two sentences**

What **is a fuzzy match**

This **is a fuzzy match**

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

1a. Similarity metrics

- **Sentence level cosine similarity (semantic). Is the cosine between the (numeric) vectors that represent two sentences**
  - **The vector representation of each word and sentence is learned by a sentence encoder (neural network), that encodes text into high-dimensional vectors**
  - **Representations learned based on aspects such as the context of words**

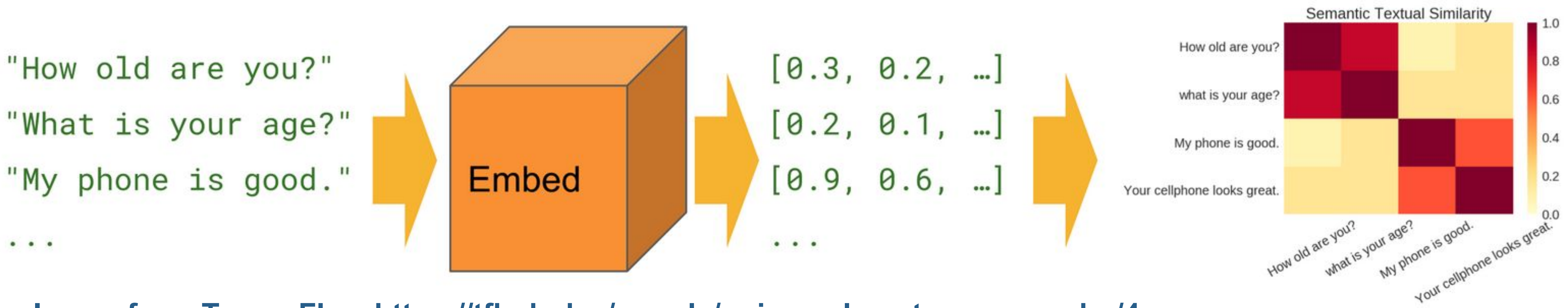
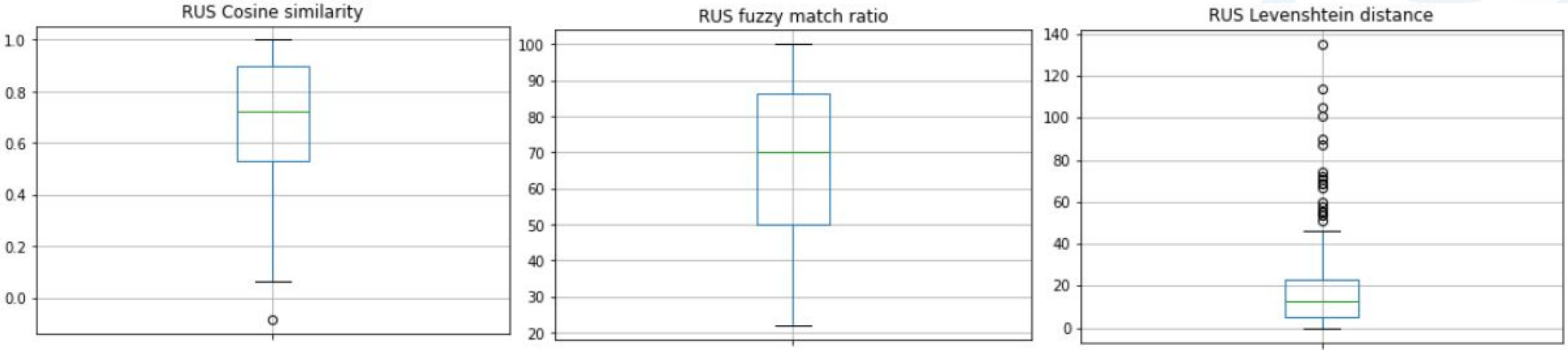


Image from TensorFlow https://tfhub.dev/google/universal-sentence-encoder/4

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

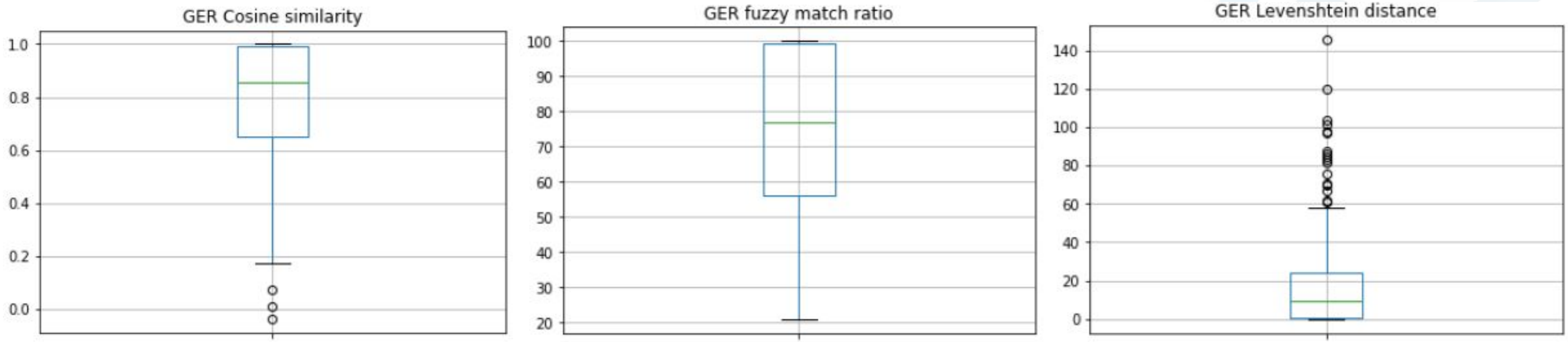
1a. visualizing similarities in RUS MT vs baseline review



- **The high cosine similarity and fuzzy match and low Levenshtein distance values indicate that the <u>MT outputs are very similar to the baseline review</u>**
  - **Cosine and Fuzzy match: the higher the better**
  - **Levenshtein distance: the lower the better**

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

1a. visualizing similarities in GER MT vs baseline review



- Again, overall the <u>MT outputs are very similar to the baseline review</u>
- Results slightly better than the Russian segments

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

1b. MT evaluation metrics

- **Bilingual Evaluation Understudy (BLEU): 2-gram weights and NIST smoothing**
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering)**
  - **Both range from 0 - 100%**
- **METEOR adds new features to BLEU, such as <u>matches based on stems and synonyms</u>**
  - **Shown to have a <u>higher correlation with human judgments</u> than BLEU for sentence level analysis**

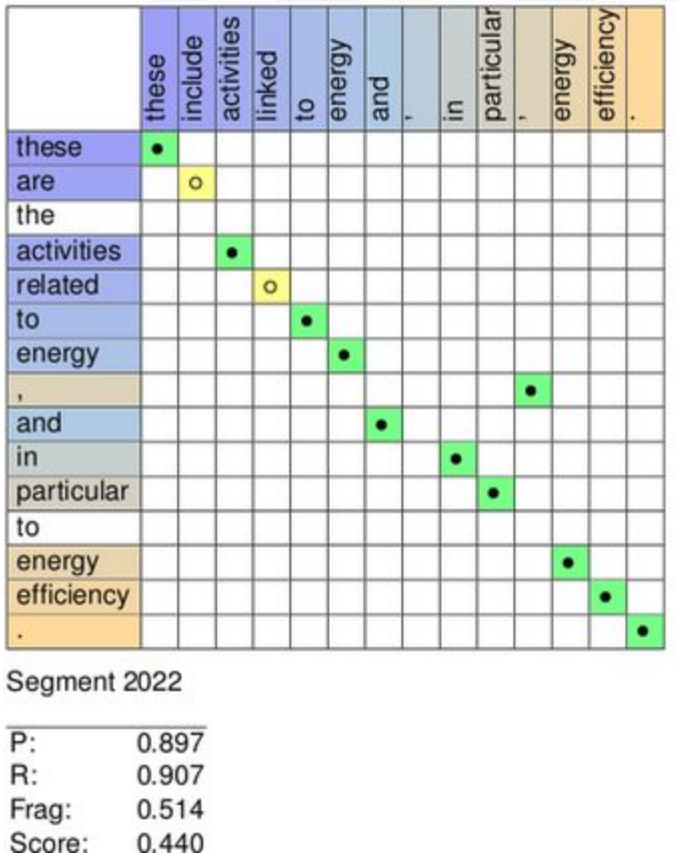


Image from https://www.cs.cmu.edu/~alavie/METEOR/examples.html

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

1b. Interpreting BLEU and METEOR scores

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

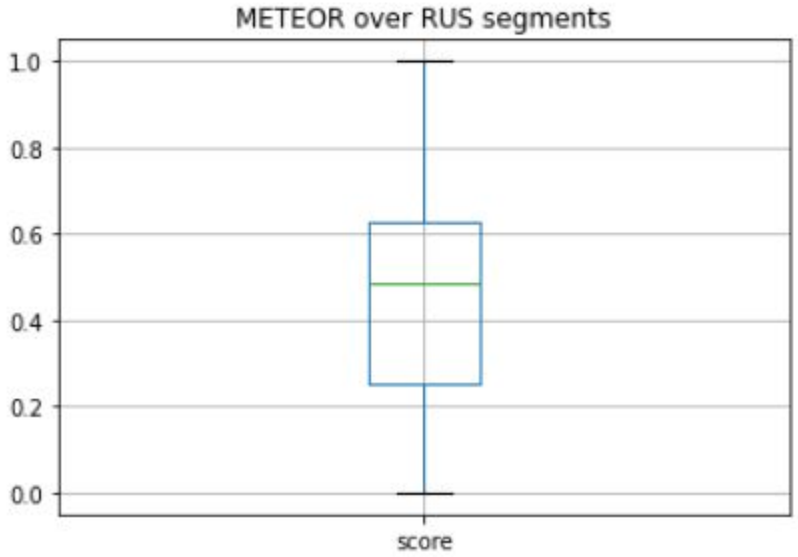
A 100% match is hard to achieve, even human translations can get around 60%-70% score due to vocabulary and phrasing differences

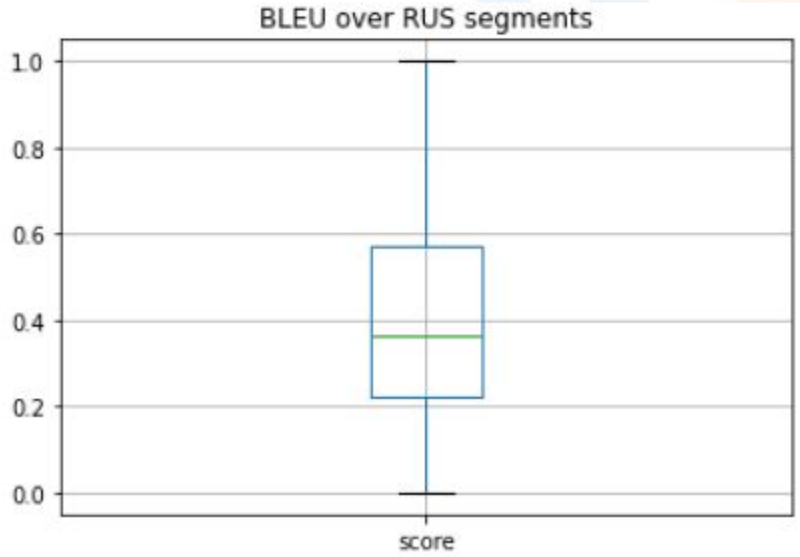
Table from https://cloud.google.com/translate/automl/docs/evaluate

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

1b. Results: MT evaluation metrics

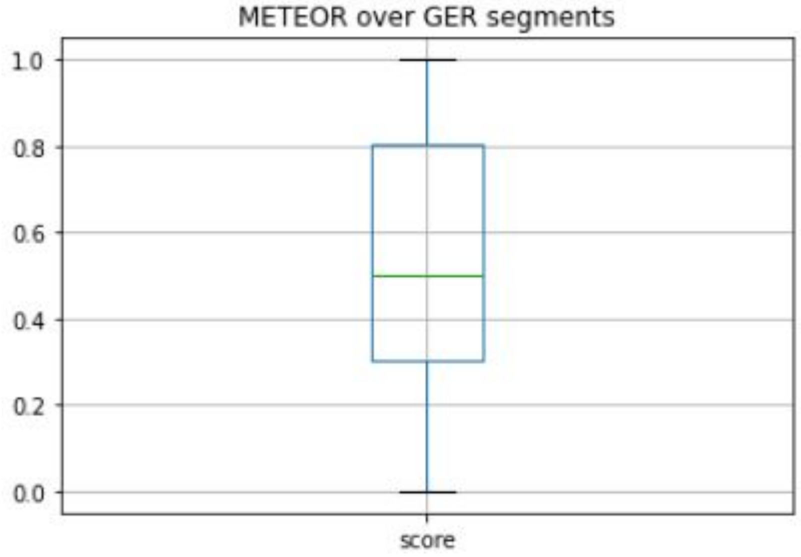


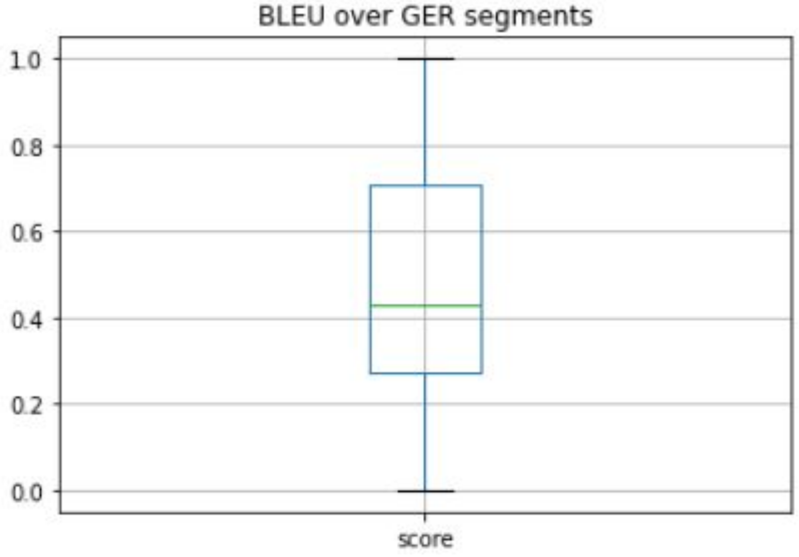


- **BLEU and METEOR metrics point that MT segments have understandable to good quality, specially when allowing synonym matches (METEOR)**
  - **Cases of really high scores probably refer to answer segments of 1 to 3 words (e.g. 'yes', 'no')**

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

1b. Results: MT evaluation metrics





- **Consistently with the similarity metrics, <u>the MT German segments have higher quality than the Russian ones</u>**
- **A median higher than 40% even for the most restrictive metric (BLEU) indicates that the MT engine produced quite good translations**

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

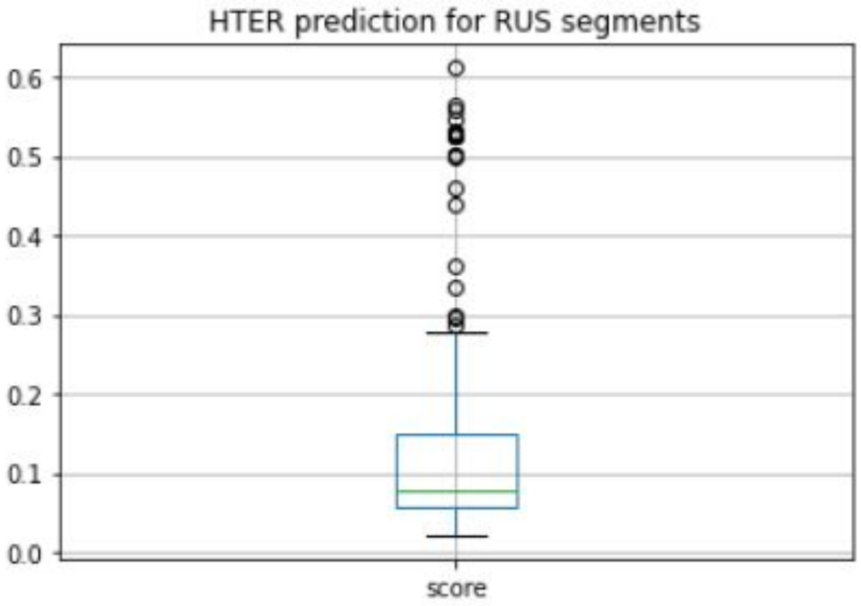
SSHOC social sciences & humanities open cloud

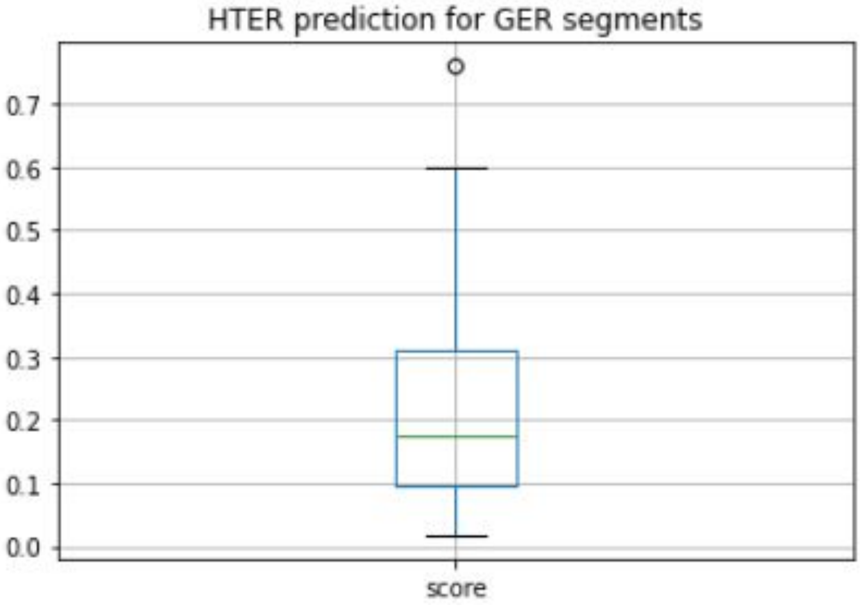
2. Quality Estimation

- Here we no longer compare the MT segments against the baseline review
  - QE models don’t need reference translations
- Sentence level Human-mediated Translation Edit Rate (HTER) prediction (the percentage of edits needed to fix the translation)
- Using TransQuest, an open source QE framework based on cross-lingual transformers
  - Data from ACL WMT19 shared task 1 (Quality Estimation)
  - Model trained with sentences in the tech domain
    - A replication of the model used by the authors in the WMT19 shared task, same hyperparameters
  - ENG-RUS: 15,089 training and 1,000 development sentences
  - ENG-GER: 13,442 training and 1,000 development sentences

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

2. HTER score predictions





- **The QE models predicted that the MT segments have good quality, overall**
  - **<u>The low HTER predictions indicate that most MT segments need very few edits to become a good quality translation</u>**
- **Lower HTER in Russian segments may indicate that the <u>models need to be fine tuned for survey domain for more reliable results</u>**

SSHOC
social sciences & humanities open cloud

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Conclusions and Future work

- **Overall the MT engine showed to produce translations sufficiently good for Post-Edition**
- **The insertion of MT+PE in the TRAPD method could minimize the human-work**
  - **Given that Quality Estimation is applied to MT segments**
- **Quality Estimation (QE) of the MT segments using a QE model trained for the survey domain**
  - **Requires post-edited data for ENG-GER, ENG-RUS**

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Thank you for your attention!

danielly.sorato@upf.edu

https://www.sshopencloud.eu

@SSHOpenCloud

info@sshopencloud.eu

/in/sshopencloud

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782