# All You Need is Source!

## A Study on Source-based Quality Estimation for NMT

**Jon Cambra** – ML Engineer
**Mara Nunziatini** – AI Program Manager

welocalize

# Agenda

Introduction

Source QE for NMT

Correlation Experiment

MTPE Prioritization Experiment

Conclusions

# Introduction

# Introduction

## Aim:

- Find a solution to **estimate the quality** of the MT output before it is produced, by looking at the **source file only** (no reference translation needed)
- Use this solution to **prioritize low-quality MT for post-editing**
- Use quality features that can be computed **easily and without training a model**

## Assumptions:

- A NMT engine will handle content **best if it is similar to the data it was trained on**
- To improve BLEU at document level, **post-editing should focus on segments that dangerously differ from the training material**

## Proposed Process:

- **Step 1: Quality Estimation**: predict the quality of raw NMT output by comparing the source content to be translated and the engine training material
- **Step 2: Post-Editing Prioritization**: error-prone segments are prioritized for post-editing by looking at the similarity scores obtained. Low-similarity segments are more likely to contain issues

# Source QE for NMT

# BOW Similarity

- Vectors describe the number of occurrences of a set of words
- With this representation we compute for each segment to translate

  the average similarity to all training segments

  the maximum similarity over all training segments

- It reduces the similarity to word matching

| | the | how | ... | is | good | day | you |
|---|---|---|---|---|---|---|---|
| It is raining. | 0 | 0 | ... | 1 | 0 | 0 | 0 |
| How are you? | 0 | 1 | ... | 0 | 0 | 0 | 1 |
| Sunny day. | 0 | 0 | ... | 0 | 0 | 1 | 0 |
| I prefer the snow. | 1 | 0 | ... | 0 | 0 | 0 | 0 |
| The kitchen is closed. | 1 | 0 | ... | 1 | 0 | 0 | 0 |

$$\mathrm{avg}_{\mathrm{bow}}(\mathrm{s}_{\mathrm{test}}) = \frac{1}{n_{\mathrm{train}}} \sum_{s \in S_{\mathrm{train}}} \mathrm{sim}(\vec{\mathrm{s}_{\mathrm{test}_{\mathrm{bow}}}}, \vec{\mathrm{s}_{\mathrm{bow}}})$$

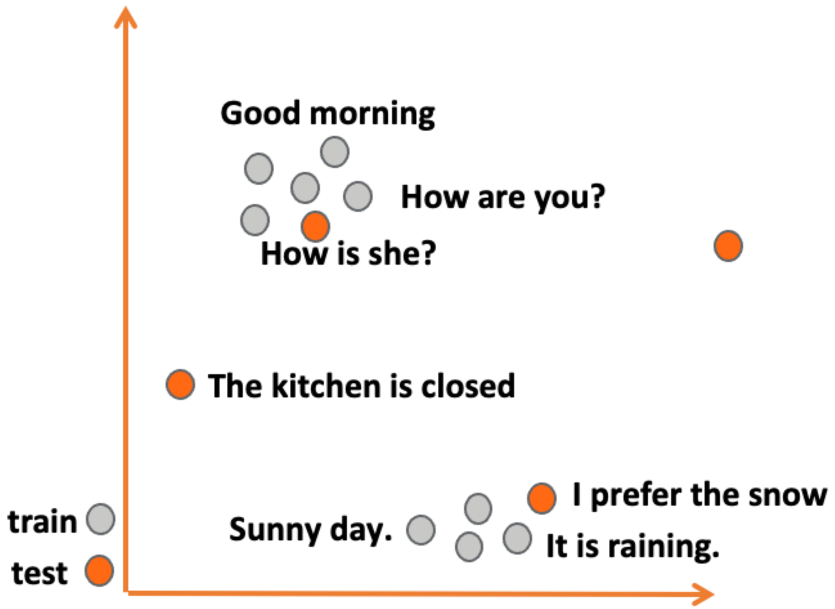
$$\mathrm{max}_{\mathrm{bow}}(\mathrm{s}_{\mathrm{test}}) = \max_{s \in S_{\mathrm{train}}} \mathrm{sim}(\vec{\mathrm{s}_{\mathrm{test}_{\mathrm{bow}}}}, \vec{\mathrm{s}_{\mathrm{bow}}})$$

# Semantic Similarity

- An accurate semantic representation can be implemented with sentence embedding models
- These models are Siamese BERT-Networks trained on SNLI data
- With this representation we compute for each segment to translate

the average similarity to all training segments

the maximum similarity over all training segments




$$\mathrm{avg}_{\mathrm{sem}}(\mathrm{s}_{\mathrm{test}}) = \frac{1}{n_{\mathrm{train}}} \sum_{s \in S_{\mathrm{train}}} \mathrm{sim}(\vec{\mathrm{s}_{\mathrm{test}}}_{\mathrm{sem}}, \vec{\mathrm{s}}_{\mathrm{sem}})$$

$$\mathrm{max}_{\mathrm{sem}}(\mathrm{s}_{\mathrm{test}}) = \max_{s \in S_{\mathrm{train}}} \mathrm{sim}(\vec{\mathrm{s}_{\mathrm{test}}}_{\mathrm{sem}}, \vec{\mathrm{s}}_{\mathrm{sem}})$$

# Unknown Words

- A segment to translate can be highly similar to a training segment but with an important difference
- If a word is not contained in the training data, the engine will probably fail
- We call this type of words: unknown words
- We propose a feature counting the number of unknown words per segment

| | |
|---|---|
| source | The best museums are in **London**. |
| hyp | Los mejores museos están en **London**. |
| ref | Los mejores museos están en Londres. |
| source | The best museums are in Madrid. |
| ref | Los mejores museos están en Madrid. |

**Table 1:** Example on NMT errors due to unknown words. The first example describes the translation produced by a NMT system. We highlight **in bold the unseen word** in training and **in red the translation error**. The second example corresponds to the most similar segment found in training with a cosine similarity of 0.95

# Correlation Experiment

# Experiment Setup

## NMT SYSTEMS

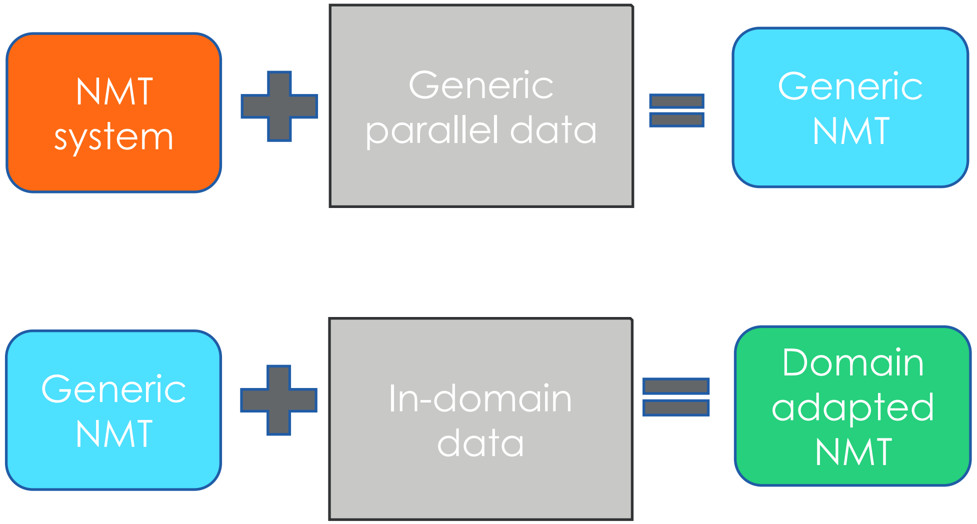
- en>de, en>it and en>ko **generic engines** are created with a large amount of data form different domains
- The generic engines are then adapted with **User Interface and User Assistance domain** data

## NMT DATA

- **Generic data**: large amounts of data extracted from OPUS
- **In-domain data**: private company TMs and glossaries

## BENCHMARK BASELINE FEATURES

- **TP**: NMT sequence-level translation probability normalized by length
- **COMET-as-QE**: score returned by a model trained on labeled data




| | Generic | In-domain |
|---|---|---|
| **En-De** | 11,568,049 | 181,061 |
| **En-It** | 32,187,643 | 89,835 |
| **En-Ko** | 17,299,009 | 173,662 |

**Table 2:** Summary table counting the amount of segment pairs used to train NMT systems

# Correlation Experiment

**How do the proposed new features correlate to the translation quality at a segment level?**

We extract the correlations between the features and different quality metrics:

- ✓ **MT SCORES** (all engines)
  - **BLEU**: commonly used token-level metric
  - **chrF3**: character-level metric
  - **COMET**: metric showing highest correlation to human DA

- ✓ **DIRECT ASSESSMENT** (customized en>it and en>de engines)
  - Adequacy + Fluency (1 is the lowest score and 5 is the highest)
  - 3 annotators
  - Scores averaged and standardized

# Results

## Generic Engines

**BLEU and chrF3**: our proposed features do not show a strong correlation

**COMET**: stronger correlation can be seen for max_sem (en>it, en>de) and max_bow (en>de)

This behavior can be explained by the nature of the translation and the limitation of the string metrics: they fail to correctly evaluate the quality of flawless translations which use different terminology or style compared to the reference.

| | En-It | | | En-De | | | En-Ko | | |
|---|---|---|---|---|---|---|---|---|---|
| | BLEU | chrF3 | COMET | BLEU | chrF3 | COMET | BLEU | chrF3 | COMET |
| TP | **0.191** | **0.376** | 0.389 | **0.200** | **0.297** | 0.423 | **0.492** | **0.662** | 0.440 |
| $\text{COMET}_{\text{QE}}$ | 0.191 | 0.166 | **0.821*** | 0.053 | 0.048 | **0.824*** | 0.048 | 0.004 | **0.622*** |
| $\text{avg}_{\text{bow}}$ | -0.077 | 0.094 | -0.099 | -0.029 | -0.063 | -0.002 | -0.021 | -0.067 | 0.037 |
| $\text{max}_{\text{bow}}$ | **0.123** | 0.093 | 0.042 | 0.006 | 0.020 | 0.168 | 0.030 | 0.041 | 0.015 |
| $\text{avg}_{\text{sem}}$ | 0.048 | **-0.133** | -0.152 | -0.129 | **-0.124** | 0.148 | 0.053 | 0.043 | **-0.198** |
| $\text{max}_{\text{sem}}$ | 0.027 | -0.063 | **0.196** | **0.132** | 0.032 | **0.324** | -0.009 | -0.050 | 0.021 |
| unk | -0.015 | -0.044 | 0.099 | -0.010 | -0.001 | -0.131 | - | - | - |

**Table 3:** Pearson correlation table between features and different automatic MT metrics for generic NMT settings. Highest and relevant correlations from all the proposed approaches are in bold; find also in bold the best result between the two baselines.

# Results

## Domain-adapted Engines

### MT SCORES

- **max_bow**: interesting correlation for en>ko only
- **max_sem**: stronger overall correlation between the proposed features
- **unk**: significant negative correlation with COMET

### DIRECT ASSESSMENT

- **max_bow**: moderate correlation
- **max_sem**: competitive correlation vs COMET as QE
- **unk**: our feature does not seem to correlate
- Highest correlation is achieved with **TP**

| | En-It | | | | | En-De | | | | | En-Ko | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | BLEU | chrF3 | COMET | Fcy | Adcy | BLEU | chrF3 | COMET | Fcy | Adcy | BLEU | chrF3 | COMET |
| **TP** | **0.230** | **0.379** | 0.349 | **0.374** | **0.456** | **0.131** | **0.336** | 0.339 | **0.217** | **0.343** | **0.344** | **0.531** | 0.379 |
| **COMET$_{QE}$** | 0.199 | 0.119 | **0.646*** | 0.326 | 0.312 | 0.102 | 0.192 | **0.604*** | 0.193 | 0.177 | 0.011 | 0.026 | **0.553*** |
| **max$_{bow}$** | 0.073 | 0.055 | 0.056 | 0.109 | 0.127 | 0.070 | 0.071 | 0.170 | 0.174 | 0.146 | **0.282** | **0.271** | 0.163 |
| **max$_{sem}$** | **0.241** | **0.161** | 0.269 | 0.246 | **0.253** | **0.264** | **0.285** | **0.355** | **0.189** | **0.175** | 0.237 | 0.224 | **0.174** |
| **unk(-)** | 0.138 | 0.078 | **0.374** | **0.282** | 0.237 | 0.139 | 0.160 | 0.333 | 0.156 | 0.072 | 0.057 | 0.065 | 0.046 |

**Table 4:** Pearson correlation between features and different automatic MT metrics and DA scores for domain adapted NMT settings. Highest correlations with all the proposed approaches are in bold; find also in bold the best result between the two baselines.

# PE Prioritization Experiment

# Settings

**en>it and en>de domain-adapted engines**

- We obtain the BLEU scores after simulating PE on a selected number of segments according to the corresponding indicators:
  - sem_max_sim

50% selection based on max similarity

| Source segment | BLEU | TP | COMET | unk | sem_max_sim |
|---|---|---|---|---|---|
| It is raining. | 45 | 0.75 | 0.53 | 0 | 0.95 |
| How are you? | 65 | 0.83 | 0.51 | 0 | 0.93 |
| Sunny day. | 85 | 0.68 | 0.47 | 0 | 0.89 |
| I prefer the snow. | 59 | 0.55 | 0.50 | 1 | 0.76 |
| The kitchen is closed. | 41 | 0.70 | 0.61 | 2 | 0.52 |
| The cat is on the mat. | 30 | 0.37 | 0.42 | 0 | 0.44 |

# Settings

**en>it and en>de domain-adapted engines**

- We obtain the BLEU scores after simulating PE on a selected number of segments according to the corresponding indicators:
  - sem_max_sim
  - unk

**50% selection based on max similarity**

| Source segment | BLEU | TP | COMET | unk | sem_max_sim |
|---|---|---|---|---|---|
| It is raining. | 45 | 0.75 | 0.53 | 0 | 0.95 |
| How are you? | 65 | 0.83 | 0.51 | 0 | 0.93 |
| Sunny day. | 85 | 0.68 | 0.47 | 0 | 0.89 |
| I prefer the snow. | 59 | 0.55 | 0.50 | 1 | 0.76 |
| The kitchen is closed. | 41 | 0.70 | 0.61 | 2 | 0.52 |
| The cat is on the mat. | 30 | 0.37 | 0.42 | 0 | 0.44 |

**33% selection based on unk**

| Source segment | BLEU | TP | COMET | unk | sem_max_sim |
|---|---|---|---|---|---|
| The kitchen is closed. | 41 | 0.70 | 0.61 | 2 | 0.52 |
| I prefer the snow. | 59 | 0.50 | 0.50 | 1 | 0.76 |
| It is raining. | 45 | 0.75 | 0.53 | 0 | 0.95 |
| How are you? | 65 | 0.83 | 0.51 | 0 | 0.93 |
| Sunny day. | 85 | 0.68 | 0.47 | 0 | 0.89 |
| The cat is on the mat. | 30 | 0.37 | 0.42 | 0 | 0.44 |

# Settings

**en>it and en>de domain-adapted engines**

- We obtain the BLEU scores after simulating PE on a selected number of segments according to the corresponding indicator:
  - sem_max_sim
  - unk
  - COMET
  - TP
  - unk+sem_max_sim: selects first segments based on unk, then segments based on sem_max_sim

**50% selection based on max similarity**

| Source segment | BLEU | TP | COMET | unk | sem_max_sim |
|---|---|---|---|---|---|
| It is raining. | 45 | 0.75 | 0.53 | 0 | 0.95 |
| How are you? | 65 | 0.83 | 0.51 | 0 | 0.93 |
| Sunny day. | 85 | 0.68 | 0.47 | 0 | 0.89 |
| I prefer the snow. | 59 | 0.55 | 0.50 | 1 | 0.76 |
| The kitchen is closed. | 41 | 0.70 | 0.61 | 2 | 0.52 |
| The cat is on the mat. | 30 | 0.37 | 0.42 | 0 | 0.44 |

**33% selection based on unk**

| Source segment | BLEU | TP | COMET | unk | sem_max_sim |
|---|---|---|---|---|---|
| The kitchen is closed. | 41 | 0.70 | 0.61 | 2 | 0.52 |
| I prefer the snow. | 59 | 0.50 | 0.50 | 1 | 0.76 |
| It is raining. | 45 | 0.75 | 0.53 | 0 | 0.95 |
| How are you? | 65 | 0.83 | 0.51 | 0 | 0.93 |
| Sunny day. | 85 | 0.68 | 0.47 | 0 | 0.89 |
| The cat is on the mat. | 30 | 0.37 | 0.42 | 0 | 0.44 |

# Settings

**en>it and en>de domain-adapted engines**

- We obtain the BLEU scores after simulating PE on a selected number of segments according to the corresponding indicator:
  - sem_max_sim
  - unk
  - COMET
  - TP
  - unk+sem_max_sim: selects first segments based on unk, then segments based on sem_max_sim
- **Benchmarks**:
  - **Random selection**: lower benchmark randomly selecting segments to PE.
  - **BLEU selection**: upper benchmark selecting segments based on the BLEU score of the translation.*

*Note: we need the reference to get this score, so it is an advantageous/unfair situation over the other features.

**50% selection based on max similarity**

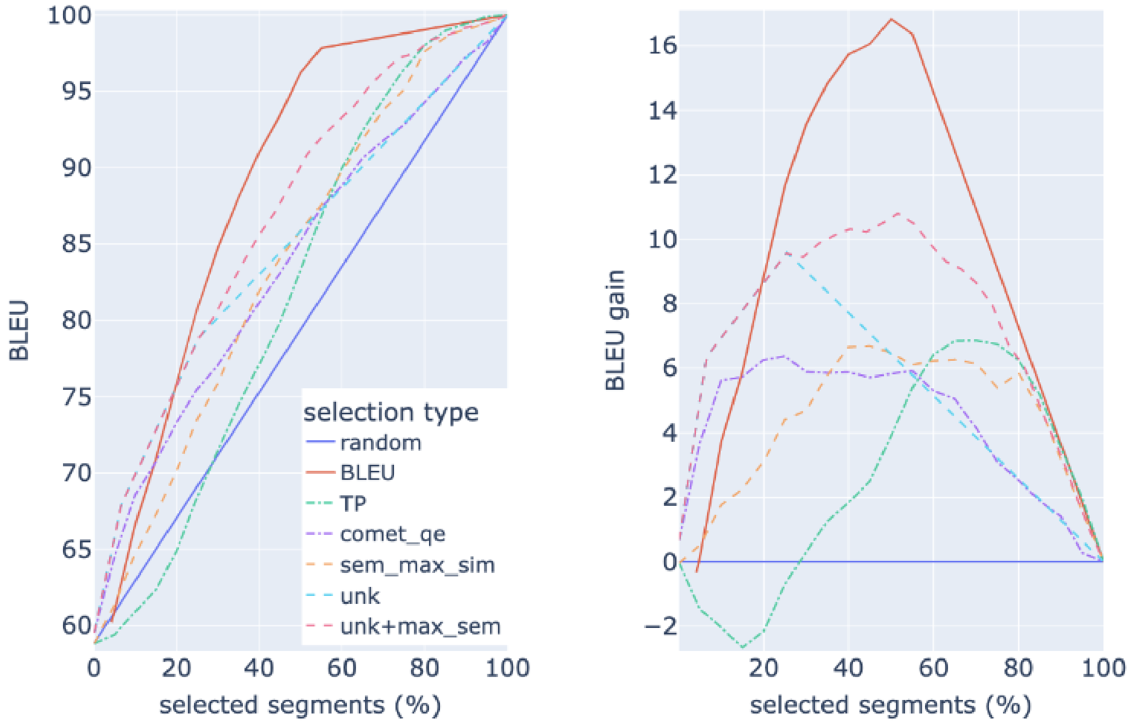
| Source segment | BLEU | TP | COMET | unk | sem_max_sim |
|---|---|---|---|---|---|
| It is raining. | 45 | 0.75 | 0.53 | 0 | 0.95 |
| How are you? | 65 | 0.83 | 0.51 | 0 | 0.93 |
| Sunny day. | 85 | 0.68 | 0.47 | 0 | 0.89 |
| I prefer the snow. | 59 | 0.55 | 0.50 | 1 | 0.76 |
| The kitchen is closed. | 41 | 0.70 | 0.61 | 2 | 0.52 |
| The cat is on the mat. | 30 | 0.37 | 0.42 | 0 | 0.44 |

**33% selection based on unk**

| Source segment | BLEU | TP | COMET | unk | sem_max_sim |
|---|---|---|---|---|---|
| The kitchen is closed. | 41 | 0.70 | 0.61 | 2 | 0.52 |
| I prefer the snow. | 59 | 0.50 | 0.50 | 1 | 0.76 |
| It is raining. | 45 | 0.75 | 0.53 | 0 | 0.95 |
| How are you? | 65 | 0.83 | 0.51 | 0 | 0.93 |
| Sunny day. | 85 | 0.68 | 0.47 | 0 | 0.89 |
| The cat is on the mat. | 30 | 0.37 | 0.42 | 0 | 0.44 |

# Results: en>de

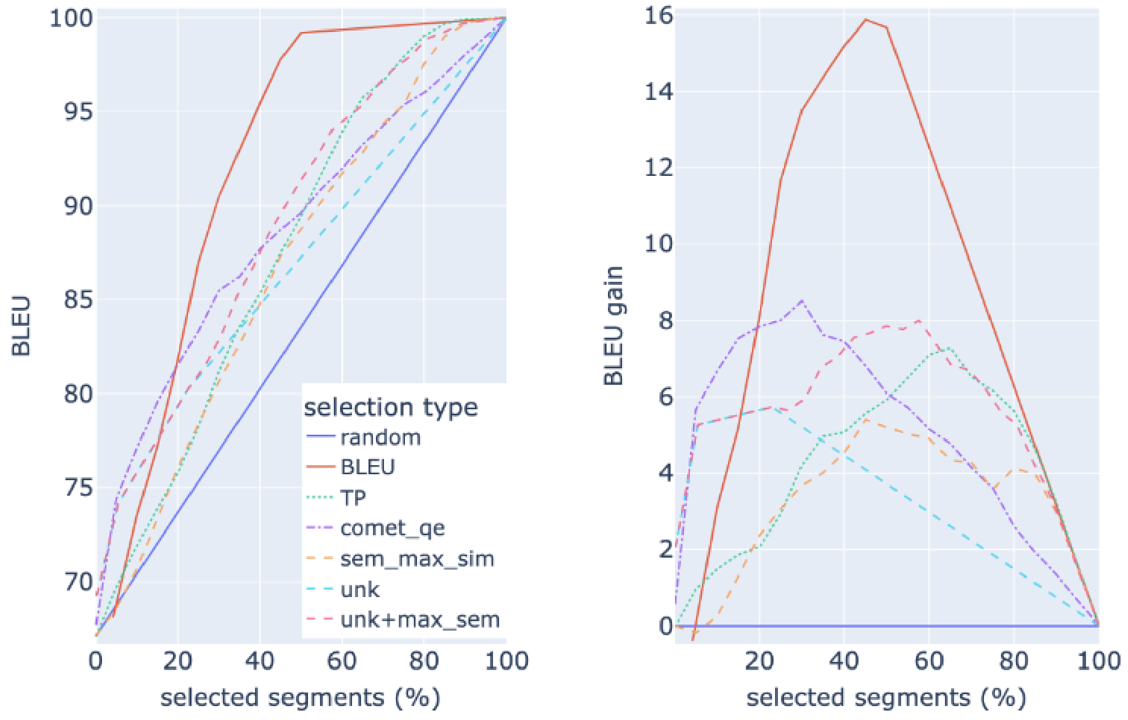
- Unk outperforms all features for the first 30%
- Sem_max_sim above 40% outperforms comet_qe and competes with TP
- Unk+max_sem takes advantage of both features and outperforms all features across all the experiment




(a) En-De BLEU evolution (b) En-De BLEU gain over random selection

# Results: en>it

- Unk is competitive for the first 20% only outperformed by comet_qe
- Sem_max_sim performs poorly on the first 40%. Above that proportion the indicator is competitive with other indicators
- Unk+max_sem takes advantage of both features and is only outperformed by comet_qe on the first 40%



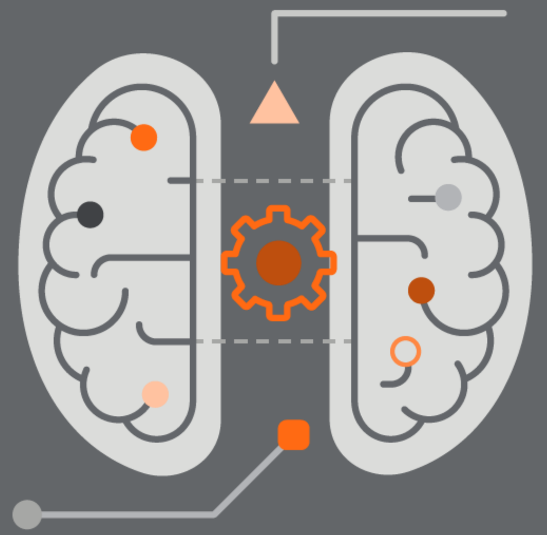

(c) En-It BLEU evolution

(d) En-It BLEU gain over random selection

# Key Takeaways

All source indicators seem to be more advantageous in some scenarios compared to Source + Translation or Source + NMT probabilities:

- unk: good indicator to select first segments while the value is superior to 1
- sem_max_sim: competitive information to decide which segments to select when you have more than 40% to post-edit
- Rule combining both indicators lead to outstanding results for every number of segments selected



# Conclusions

# Conclusions

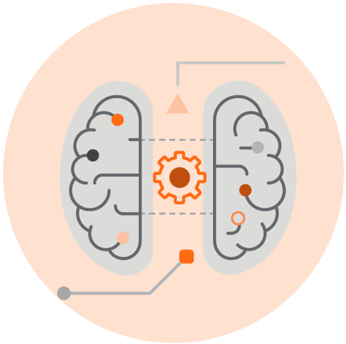
## Correlation experiment

- **The proposed features provide information about the quality of raw MT output and do not need any reference translation**
    - Generic engines: no strong correlation with string metrics, but strong correlation with COMET
    - Domain-adapted engines: strong correlation with MT metrics and DA

## MTPE prioritization


- **unk** feature is a good indicator to **initially** select challenging segments that need PE
- It is beneficial to use the **sem_max_sim** feature to prioritize segments for PE when you have **more than 40%** of the file to post-edit
- **Combining both features is the preferred solution** because it benefits from both unk and sem_max_sim

## Looking to the future



- **Engine update**: select challenging segments, perform MTPE on these segments and add post-edited segments to the engine training material to improve the engine's performance
- Future QE models should make use of features that consider the **similarity or domain shift** between translation data and training data

# Thank you

welocalize

# And Special Thanks To:

Anna Pizzolato
David Clarke
Elaine O'Curran
Lena Marg
Matthew Dixon

welocalize