APE-QuEst

Connecting Europe Facility 2014-2020
CEF TELECOM
CEF-TC-2017-3, 2017-EU-IA-0151

2018-2020

Validating Quality Estimation in a CAT Workflow: Speed, Cost and Quality Trade-off

Fernando Alva-Manchego, Lucia Specia, Sara Szoc, Tom Vanallemeersch, Heidi Depraetere

CROSSLANG TRANSLATION AUTOMATION

The University Of Sheffield.

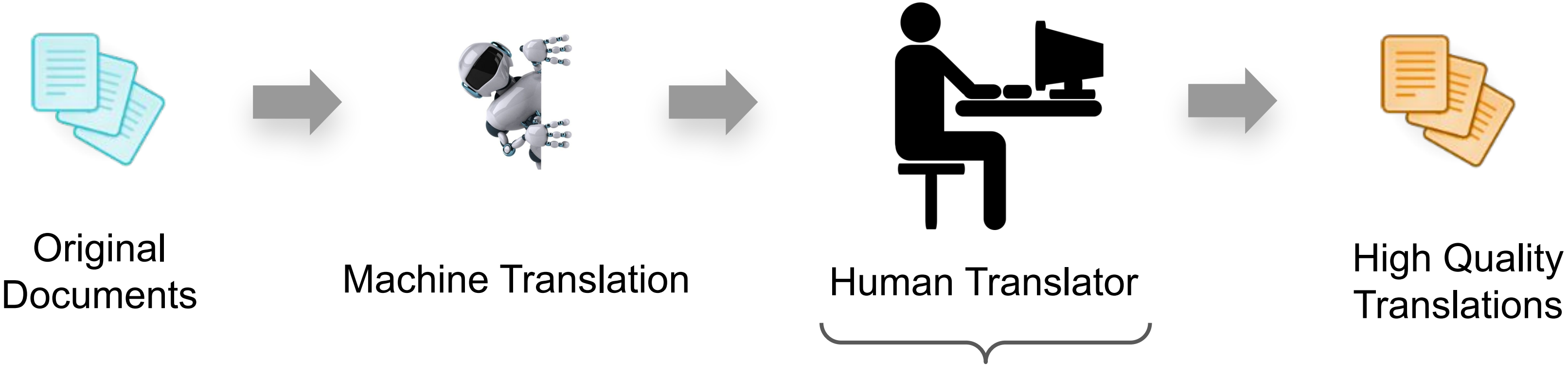
Unbabel

MT Summit 2021

Outline

- APE-QuEst: A Quality Gate for Machine Translation
- Evaluation Protocol
- Experiments and Results
- Conclusions

CAT Workflow with Machine Translation



Original Documents

Machine Translation

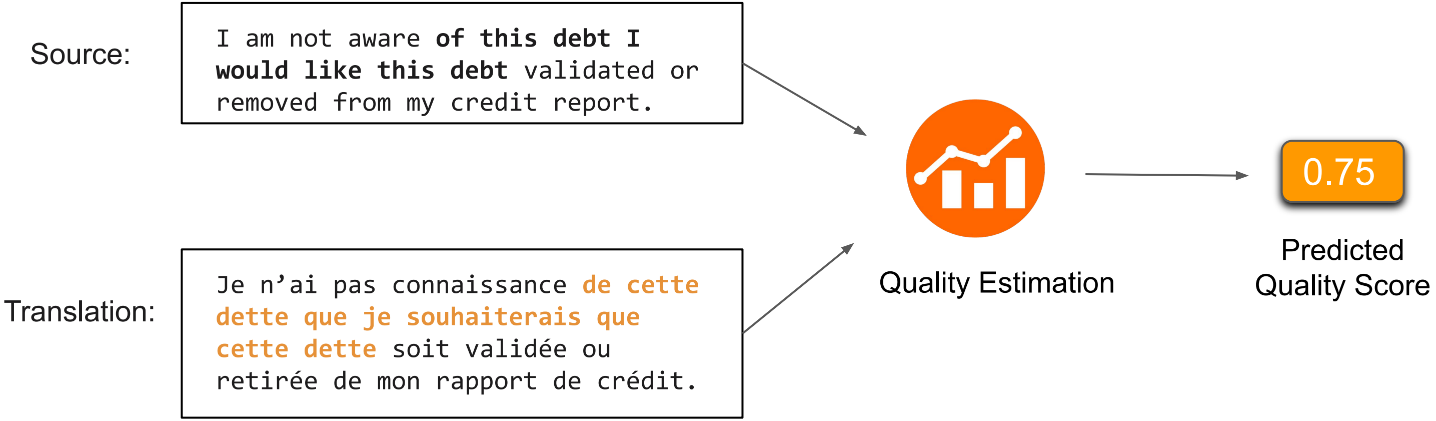
Human Translator

High Quality Translations

MT outputs are manually checked and edited as needed

How to speed up this process taking advantage of acceptable quality MT outputs?

Strategy: Automatic Quality Estimation



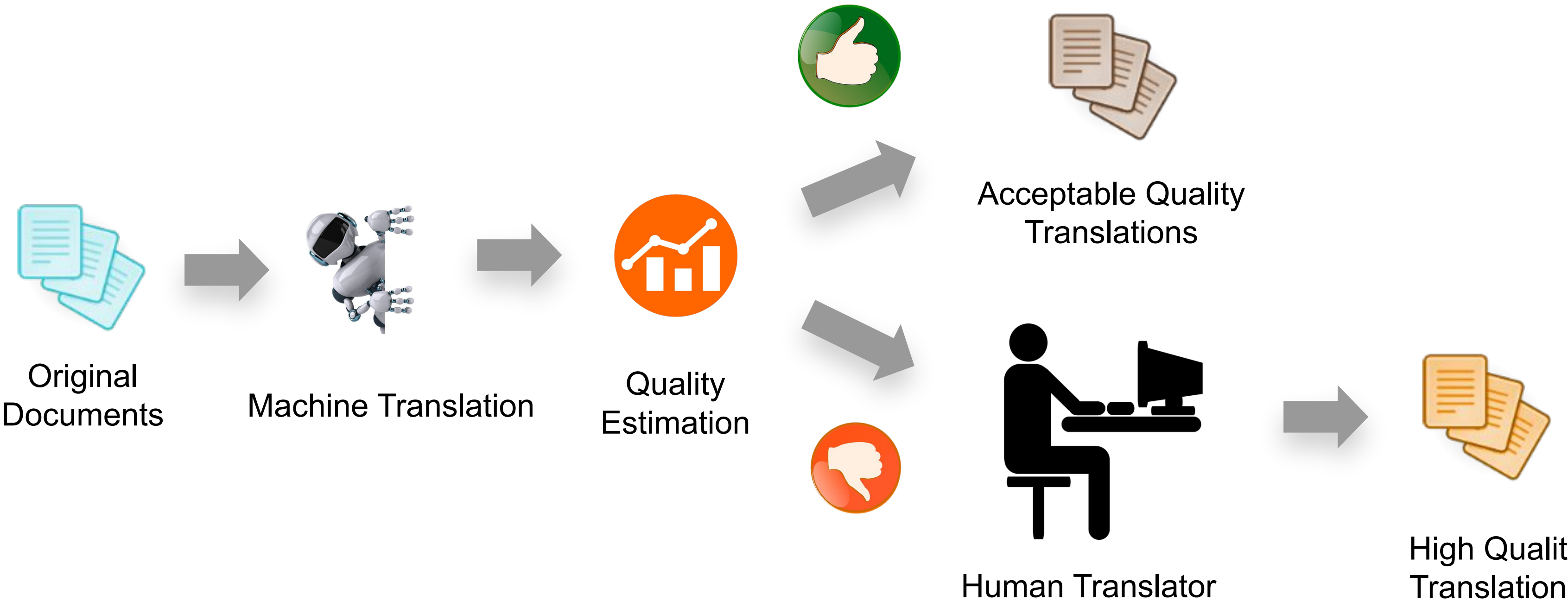
Source:

I am not aware **of this debt I would like this debt** validated or removed from my credit report.

Translation:

Je n’ai pas connaissance **de cette dette que je souhaiterais que cette dette** soit validée ou retirée de mon rapport de crédit.

CAT Workflow with MT + QE



Original Documents

Machine Translation

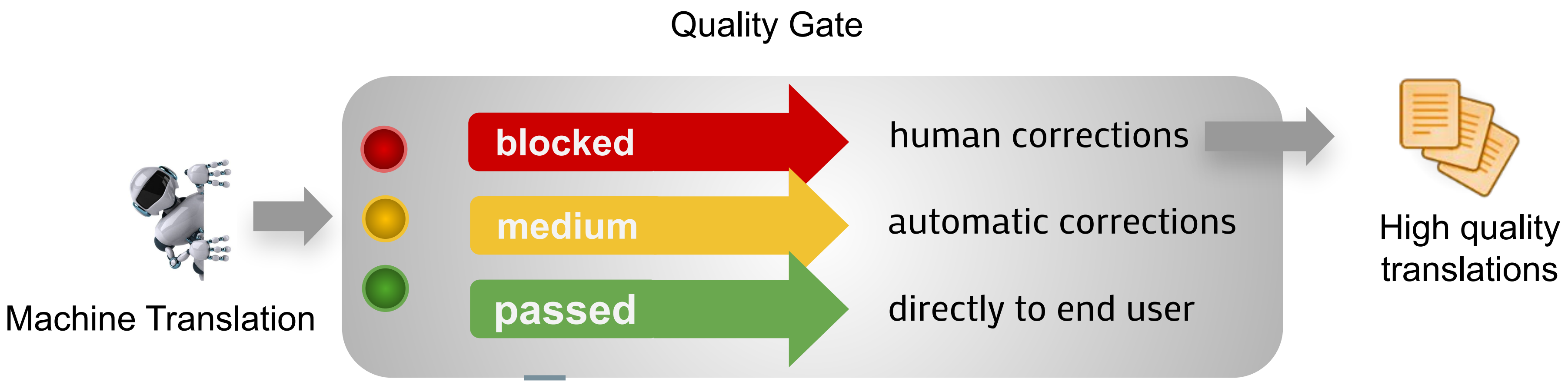
Quality Estimation

Acceptable Quality Translations

Human Translator

High Quality Translations

A Quality Gate for Machine Translation



Quality Gate

Machine Translation

blocked — human corrections

medium — automatic corrections

passed — directly to end user

High quality translations

MT output is scored and sent to 3 separate workflows

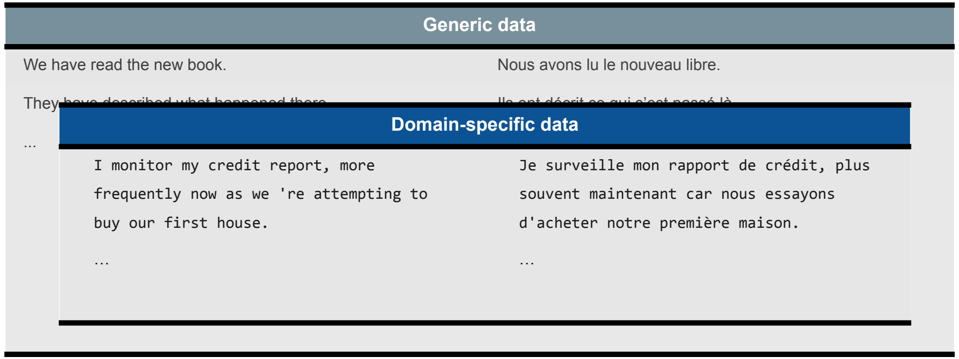
How to determine QE scores thresholds for each workflow?

How do these thresholds affect speed/time/cost?

Experimental Setup: Machine Translation

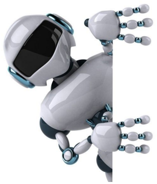
- eTranslation
  - Neural MT models for more than 24 languages
  - Targeted mainly at European public administrations

- Experiments include English → Dutch, English → French



Generic data	
We have read the new book.	Nous avons lu le nouveau libre.
They have described what happened there.	Ils ont décrit ce qui s'est passé là.
...	

Domain-specific data	
I monitor my credit report, more frequently now as we 're attempting to buy our first house.	Je surveille mon rapport de crédit, plus souvent maintenant car nous essayons d'acheter notre première maison.
...	...



Experimental Setup: Quality Estimation

- Translation Quality Estimation at the **sentence level**

- Prediction of quality score: 1 - HTER score
  - HTER: Human Translation Edit Rate

1 = Everything was changed

0 = Nothing was changed

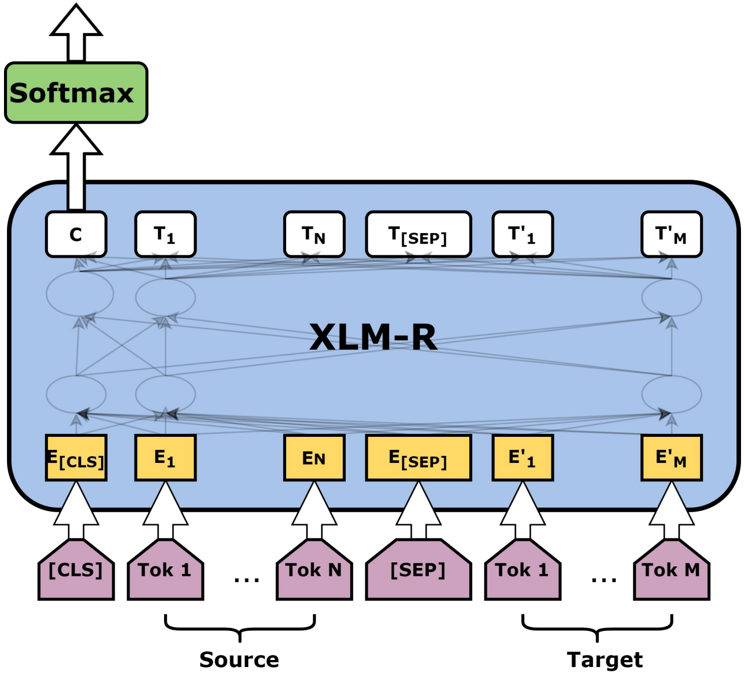
OpenKiwi
By Unbabel

github.com/Unbabel/OpenKiwi

TransQuest

github.com/TharinduDR/TransQuest

Experimental Setup: Quality Estimation



- Trained language-specific models by fine-tuning Multilingual BERT
- Training dataset from Ive et al. (2020)
  - Tuples: (source, MT, human post-edition, target)
  - Legal domain
  - Size:
    - 11,249 for English-Dutch (EN-NL)
    - 9,989 for English-French (EN-FR)

	EN-NL		EN-FR	
Model	r	MAE	r	MAE
Ive et al. (2020)	0.38	0.14	0.58	0.14
Ours	**0.51**	**0.10**	**0.69**	**0.10**

Research Questions

- When compared to the **Traditional** workflow, **does the Quality Gate workflow help to improve speed** (i.e. time to get to final translation) and **reduce cost** (how many translations need HPE)?
- When compared to the **MT-Only** workflow, **does the Quality Gate workflow help to improve translation quality**?

Evaluation Protocol: Measurable Criteria

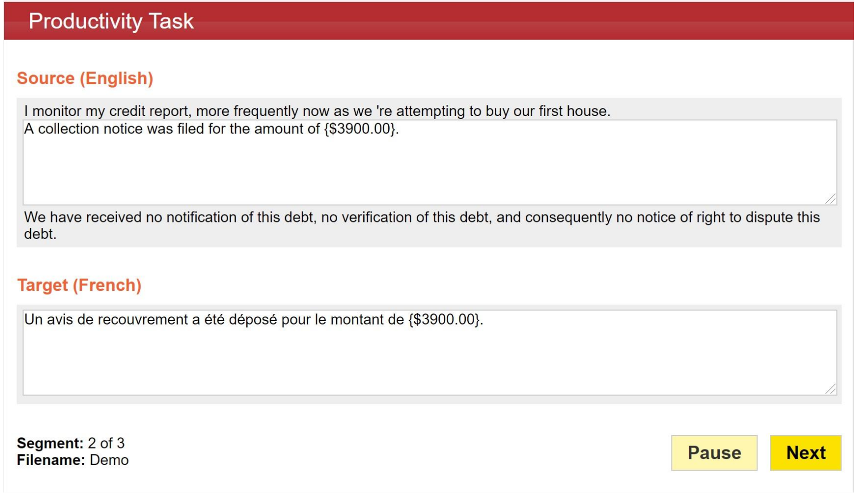
- **Quality:** Percentage of sentences considered of acceptable quality by independent human raters
- **Cost:** Percentage of sentences that require human post-editing, versus being fit for purpose
- **Speed:** Time required for human post-edition

Evaluation Protocol: Use Cases

- Texts sampled from a European public administration handling consumer complaints

Use Case 1 (Assimilation)	Use Case 2 (Dissemination)
Consumer Complaints (informal)	Privacy Statement (formal)
Content needs to be **understood**	Content needs to be **published**
Is the translation Acceptable / Not Acceptable ?	Is the translation Publishable / Not Publishable ?
966 English source sentences	114 English source sentences

Data Collection: Human Post-Edits*



Source (English)

I monitor my credit report, more frequently now as we 're attempting to buy our first house.

A collection notice was filed for the amount of {$3900.00}.

We have received no notification of this debt, no verification of this debt, and consequently no notice of right to dispute this debt.

Target (French)

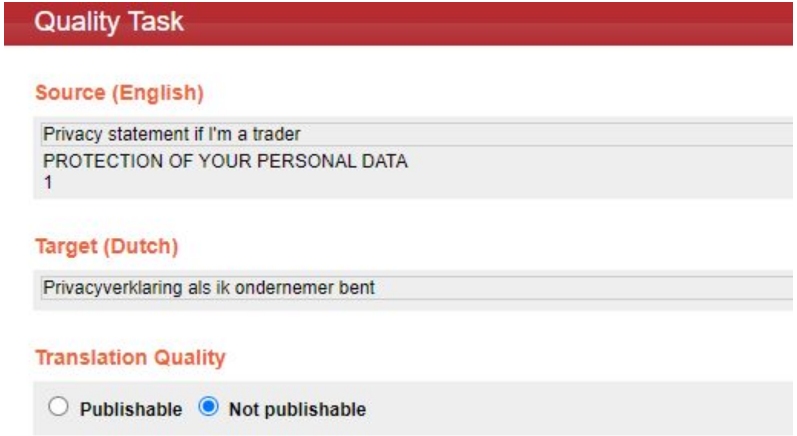
Un avis de recouvrement a été déposé pour le montant de {$3900.00}.

Segment: 2 of 3
Filename: Demo

- All MT outputs were post-edited (in each use case and target language)
- Post-editors were experienced professional translators
- For each target language, three post-editors were hired, and each sentence was post-edited once
- Sentences were post-edited within their document context

*For confidentiality reasons, the example originates from a comparable, publicly available dataset (US Government)

Data Collection: Acceptability Ratings



Quality Task

Source (English)

Privacy statement if I'm a trader
PROTECTION OF YOUR PERSONAL DATA
1

Target (Dutch)

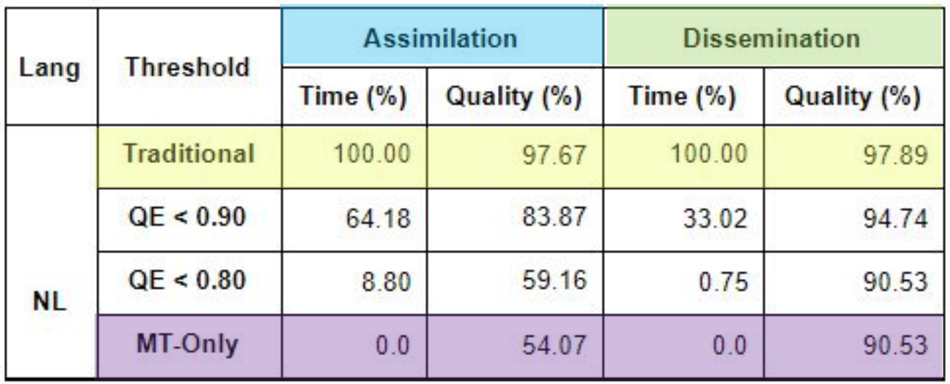
Privacyverklaring als ik ondernemer bent

Translation Quality

○ Publishable ● Not publishable

- All MT outputs and HPEs were rated
- Raters were professional translators
  - They were not informed of whether the sentences being judged were an MT output or HPE
- For each target language and use case, two raters scored each translation (either MT or HPE) once
- Sentences were rated within their document context

Results: Predicted Scores

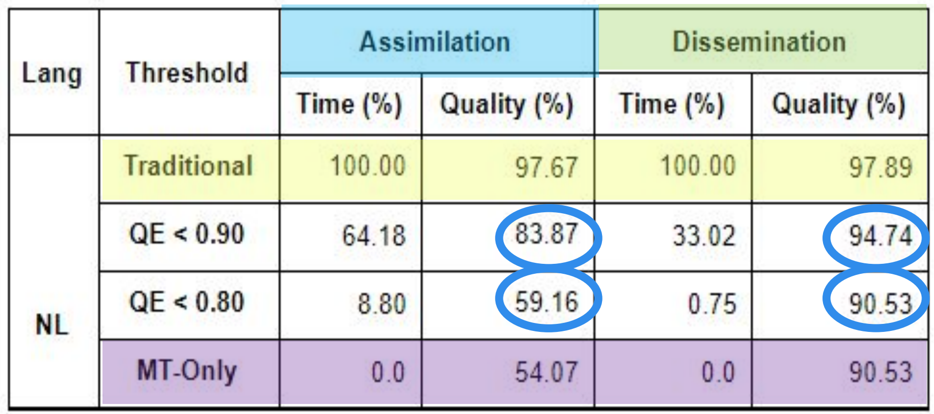


Lang	Threshold	Assimilation		Dissemination	
		Time (%)	Quality (%)	Time (%)	Quality (%)
NL	Traditional	100.00	97.67	100.00	97.89
	QE < 0.90	64.18	83.87	33.02	94.74
	QE < 0.80	8.80	59.16	0.75	90.53
	MT-Only	0.0	54.07	0.0	90.53

- High quality gains compared to MT only workflow
- Cost / time savings compared to traditional workflow
- Less impact on dissemination dataset due to high quality in-domain MT output

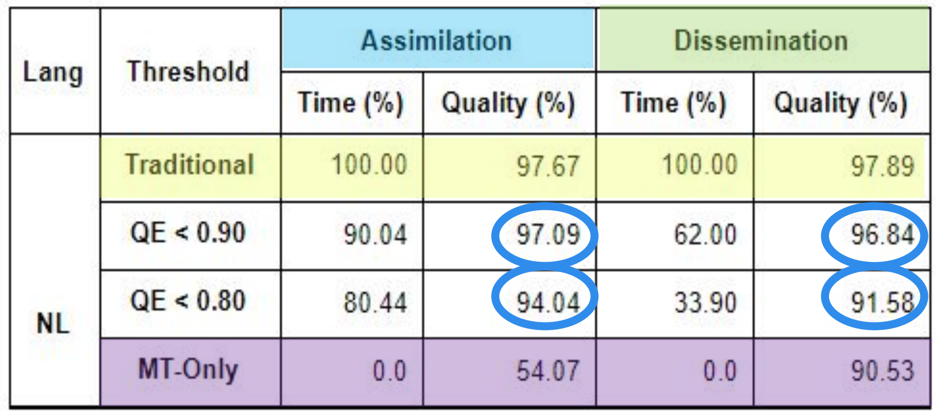
Results: Predicted vs Oracle Scores

Predicted



Lang	Threshold	Assimilation		Dissemination	
		Time (%)	Quality (%)	Time (%)	Quality (%)
NL	Traditional	100.00	97.67	100.00	97.89
	QE < 0.90	64.18	83.87	33.02	94.74
	QE < 0.80	8.80	59.16	0.75	90.53
	MT-Only	0.0	54.07	0.0	90.53

Oracle



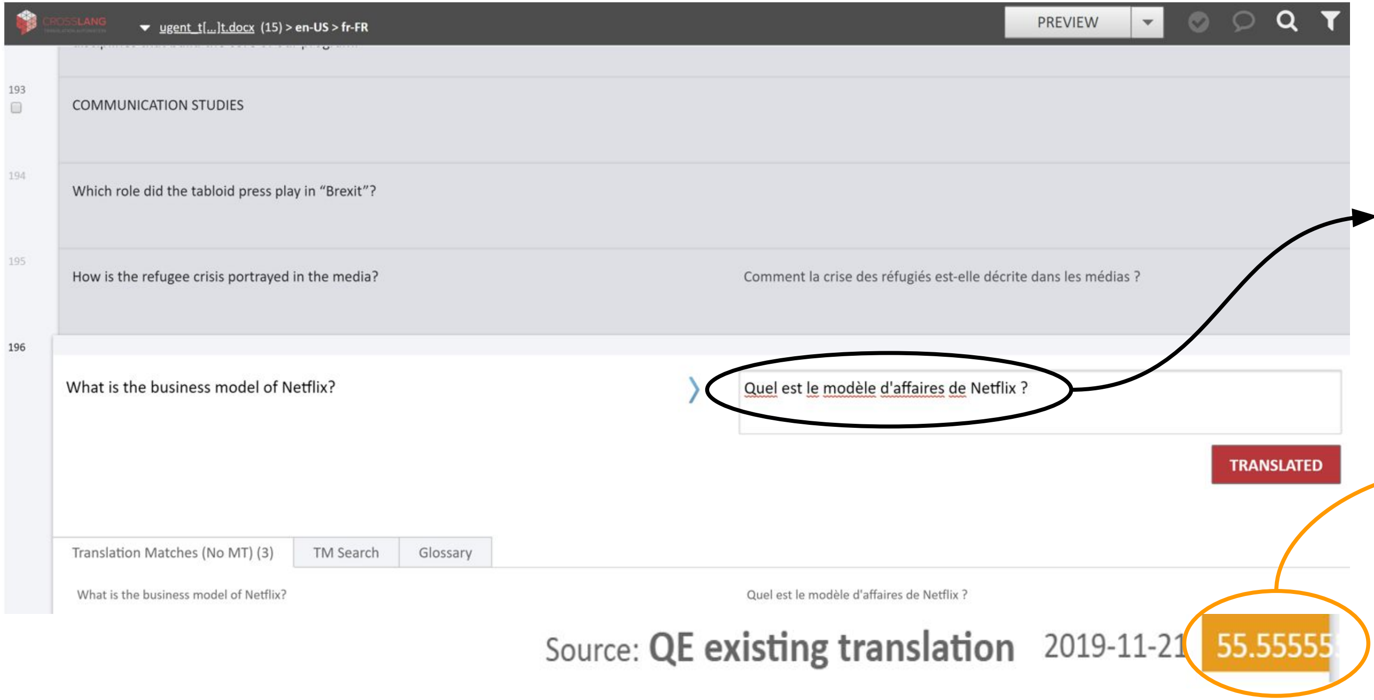
Lang	Threshold	Assimilation		Dissemination	
		Time (%)	Quality (%)	Time (%)	Quality (%)
NL	Traditional	100.00	97.67	100.00	97.89
	QE < 0.90	90.04	97.09	62.00	96.84
	QE < 0.80	80.44	94.04	33.90	91.58
	MT-Only	0.0	54.07	0.0	90.53

- Oracle scores show that there is still much room for improving QE predictions

Conclusions

- Evidence of the benefits of introducing QE into a CAT workflow
- **APE-QuEst Quality Gate:** use QE scores to determine if MT outputs can be used as-is (acceptable quality) or if they require post-edition (unacceptable quality)
- **Trade-off study:** establish thresholds on the QE scores
    - We collected human post-edits and acceptability ratings from real use case scenarios
    - Quality Gate can **obtain similar levels of quality** to the current human-only workflow, for all use cases and target languages explored

Interface to the Quality Gate



CROSSLANG ugent_t[...]t.docx (15) > en-US > fr-FR

193 COMMUNICATION STUDIES

194 Which role did the tabloid press play in “Brexit”?

195 How is the refugee crisis portrayed in the media? — Comment la crise des réfugiés est-elle décrite dans les médias ?

196 What is the business model of Netflix? — Quel est le modèle d'affaires de Netflix ?

What is the business model of Netflix? — Quel est le modèle d'affaires de Netflix ?

Source: QE existing translation 2019-11-21 55.55555

MT output /
Human post-edition /
APE output

QE score

Thank you!

APE-QuEst

@ape_quest

www.ape-quest.eu

CROSSLANG
TRANSLATION AUTOMATION

The University Of Sheffield.

Unbabel