MTET Evaluation for MT Engines

MT Evaluator Instructions

Version 2.1, October 2020

Document Owner:	PANGEANIC BI EUROPA SL

Version

Version	Date	Description	Author
1.0	5-05-2020	MT Evaluator Instructions	Maria Angeles Escrivà
2.0	15-10-2020	MT Evaluator Instructions new format and review	Amando Estela
2.1	28-10-2020	Bibliography	Maria Angeles Escrivà

Table of Contents

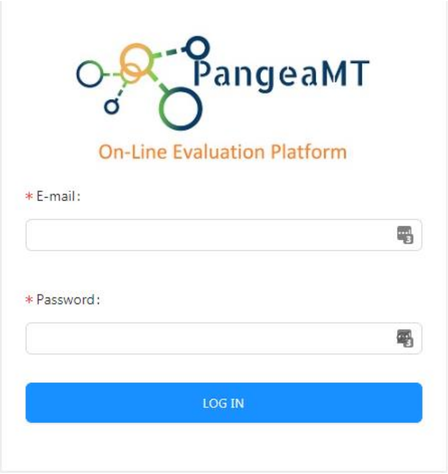
1.How to start

Welcome evaluator!

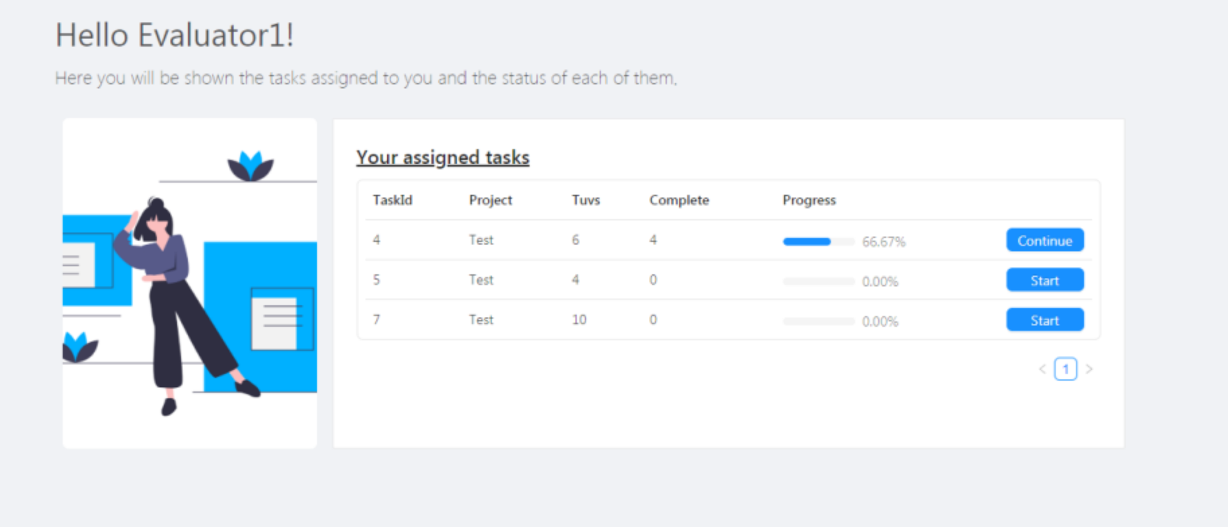
MTET is the official tool of Pangeanic's NTEU project for EU government machine translation testing which has been customized for commercial use. It will serve as a working environment to carry out all evaluations.

To access the **MTET** system, click on the following link:
https://mtet2.pangeamt.com

The following screen will appear. Enter your credentials.



The app will greet you with the following dashboard.



Hello Evaluator1!

Here you will be shown the tasks assigned to you and the status of each of them.

Your assigned tasks

TaskId	Project	Tuvs	Complete	Progress	
4	Test	6	4	66.67%	Continue
5	Test	4	0	0.00%	Start
7	Test	10	0	0.00%	Start

Fig. 1 Evaluator Welcome Screen

This dashboard shows the task(s) assigned to you (see some examples in the image in 2. Evaluation).

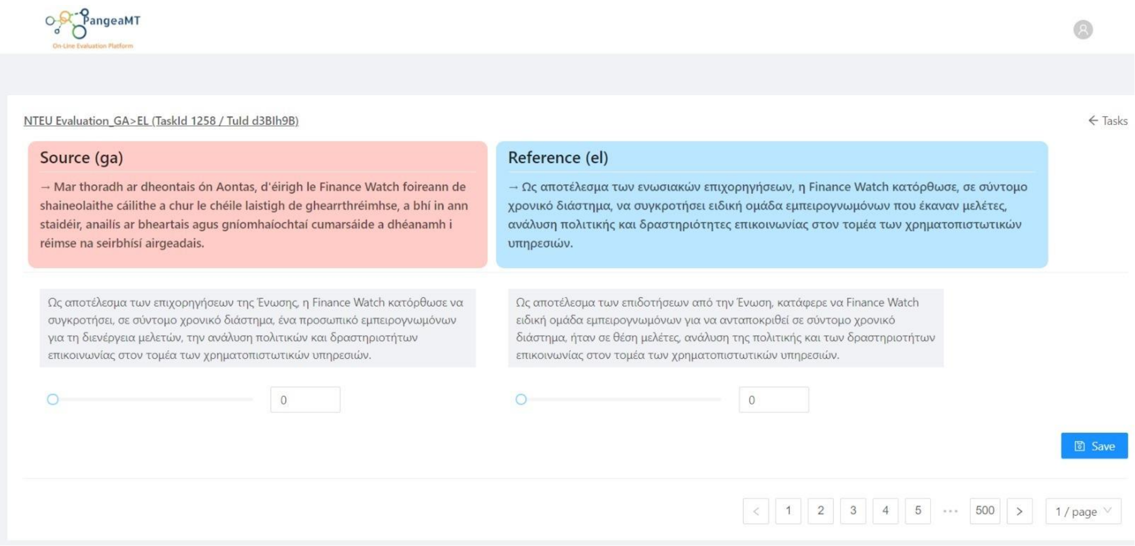
Each task includes reviewing a different language pair (eg. Spanish into French).

2. Evaluation

Click on "Start" to begin the evaluation of a language pair.

It is possible to leave the evaluation unfinished (if you need to stop and come back later, see Fig 2). However, segment evaluation being done consecutively, one sentence after another is preferable.

When accessing an unfinished task, instead of the "Start" button, the "Continue" button will be available. When you click on "Continue", the application will take you to the last phrase you were working on.



NTEU Evaluation_GA>EL (TaskId 1258 / TuId d3BIh9B)

Source (ga)

→ Mar thoradh ar dheontais ón Aontas, d'éirigh le Finance Watch foireann de shaineolaithe cáilithe a chur le chéile laistigh de ghearrthréimhse, a bhí in ann staidéir, anailís ar bheartais agus gníomhaíochtaí cumarsáide a dhéanamh i réimse na seirbhísí airgeadais.

Reference (el)

→ Ως αποτέλεσμα των ενωσιακών επιχορηγήσεων, η Finance Watch κατόρθωσε, σε σύντομο χρονικό διάστημα, να συγκροτήσει ειδική ομάδα εμπειρογνωμόνων που έκαναν μελέτες, ανάλυση πολιτικής και δραστηριότητες επικοινωνίας στον τομέα των χρηματοπιστωτικών υπηρεσιών.

Ως αποτέλεσμα των επιχορηγήσεων της Ένωσης, η Finance Watch κατόρθωσε να συγκροτήσει, σε σύντομο χρονικό διάστημα, ένα προσωπικό εμπειρογνωμόνων για τη διενέργεια μελετών, την ανάλυση πολιτικών και δραστηριοτήτων επικοινωνίας στον τομέα των χρηματοπιστωτικών υπηρεσιών.

○ 0

Ως αποτέλεσμα των επιδοτήσεων από την Ένωση, κατάφερε να Finance Watch ειδική ομάδα εμπειρογνωμόνων για να ανταποκριθεί σε σύντομο χρονικό διάστημα, ήταν σε θέση μελέτες, ανάλυση της πολιτικής και των δραστηριοτήτων επικοινωνίας στον τομέα των χρηματοπιστωτικών υπηρεσιών.

○ 0

Fig. 2 Typical Evaluation Screen

The screen above is the environment you will be working in most of the time. You will find the following fields:

- The first line at the top left is the name of the task (eg. Test (TaskId 4 / TuId Inymbll)).
- The button ← Tasks is used to return to the previous screen (dashboard).
- In the "**Source (xx)**" section you will find the **sentence that has been translated by the MT engine**.
- In the "**Reference (xx)**" section you will find the human [gold standard] reference in the target language
- Under the previous fields you will find the translation or translations to be evaluated (in NTEU we provide a second translation by a MT provider to be used as benchmark but test is blind and evaluators ignore which translation is NTEU and which is the benchmark MT provider). Below the translation, there is a slider that moves from right to left. This is the

evaluation bar.

To enter your evaluation, you can move the bar from left to right or type the number in the numerical field on the right.

- The evaluation is based from 0 to 100.
- During the evaluation, you must assess whether the machine-generated sentence adequately expresses the meaning contained in the source, that is, how close it is to how a human would have written it.

3. Evaluation Criteria

Different people may have different linguistic preferences which can affect sentence evaluation. Thus, it is important to follow the same scoring guidelines.

To standardize criteria, we will use proven academic methods to guarantee all evaluators follow the same scoring methods across languages.

Unlike SMT methods (based on bleu score) NMT needs to be ranked on accuracy, fluency and terminology.

Accuracy is defined as a sentence containing the meaning of the original, even though synonyms may have been used.

Fluency is the grammatical correctness of the sentence (gender agreements, plural / singular, case declension, etc.)

Adequacy [Terminology] is the proper use of in-domain terms agreed by the client and the developer and that are for use in production but may not be standard or general terms (the specific jargon).

When ranking a sentence, please bear the following in mind:

- **Accuracy : 33%**
- **Fluency : 33%**
- **Adequacy [terminology] : 33%**

In general, we will evaluate from 5 to 10 points for every serious error. Your evaluation must be the result of applying these discounts.

For instance, if you find two accuracy errors in the sentence (some information is missing and non-related additional information has been added), you may subtract 5% for the small error and 20% for the serious error from the Accuracy total. If you additionally find a small fluency error, you can decide additionally deducting -5%, too. Finally, if this same sentence does not contain any terminology issues, nothing is subtracted from the Adequacy percentage.

In the following example, let's assume an English original and a translation (also in English for clarity):

Original

The purpose of the ABC classification is that the effort saved in controlling and recording the resources of Group C should be directed to items of greater importance for the client or, otherwise, more important from the point of view of relevance to the organisation.

Translation

The purpose of the ABC classification is that the effort saved in the control and **registration** of Group C resources is directed to *elements* of greater importance to the *customer* or, if not, more important from the point of view of relevance to the organisation.

In the above, we can say that Accuracy is high because the meaning of the original is well expressed in the translation, although there is a term (registration instead of recording) that can be classed as incorrect. We will deduct 5% as it is a term easily corrected. Fluency is good in English, no points to deduct. In Adequacy, we note 2 small variations (items/elements, not really relevant) and client/customer. Deducting 0, 1, 2, 3, 4, 5, 6, 7 or 8% would be subjective in this case.

- When you finish scoring the sentence, click on the button Save and you will go to the next sentence.

To exit the application, click on the button 👤 appearing on the main screen and choose the "Sign out" option.

References

Measuring Machine Translation Quality in the Era of Neural: https://slator.com/academia/measuring-machine-translation-quality-in-the-era-of-neural/

What Level of Quality can Neural Machine Translation Attain on Literary Text? https://arxiv.org/pdf/1801.04962.pdf

Evaluating Explanation Methods for Neural Machine Translation

https://www.aclweb.org/anthology/2020.acl-main.35.pdf

Statistical Machine Translation Draft of Chapter 13: Neural Machine Translation

https://arxiv.org/pdf/1709.07809.pdf