

MT Quality Assessment Guidelines

The goal of this task is to evaluate the quality of Machine Translation (MT). Given a source sentence and the MT, the annotators must **rate the overall quality** of the translated sentence and **highlight the words that justify their rating**. Annotation must be conducted **independently** by each annotator.

The annotation will be conducted using the Appraise annotation environment. Each annotator will be given an individual unique link to access their annotation task.

You will evaluate 1,000 sentences randomly extracted from Wikipedia and translated into English by a state-of-the-art neural MT system. For each sentence we provide the link to the corresponding Wikipedia article. **This link must be consulted in case it is not clear what the original sentence is talking about.**

For each pair of sentences you will have the following information:

- Original sentence
- Machine translation
- Translation context: a link to the Wikipedia article

Appraise Dashboard esteng0104

0/100 blocks, 7 items left in block Explainable Quality Estimation: EST - ENG

Tundub , et sellel kaalutlusel on eriline jõud Jimi juhtumil .
— Source text - Click on the words that explain your assessment of translation quality

It seems that this consideration has a special force in the case of Jim .
— Machine translation - Click on the words that explain your assessment of translation quality

https://et.wikipedia.org/wiki/Utilitarianism:_For_and_Against
— Wikipedia article with the context of the translation

Not evaluated
— Evaluate the quality of the translation using the scale from Completely Incorrect to Perfect Translation.

Reset Submit

Sentence-level evaluation

First, you need to rate the quality of the MT on a scale from low (left) to high (right). The following hints for the interpretation of the ranges of the quality scale will appear as you move the slider.

- If the **MT output is considered perfect**, no words need to be highlighted.
- Otherwise, you must **highlight the words corresponding to the errors in translation** that made you assign an imperfect score. The **corresponding source words that caused the errors must also be highlighted**.
- **All and only** the words necessary to justify the sentence score must be highlighted.
- **If some content is missing from the MT output, highlight the source words that were not translated.**
- If the MT output has some additional contents, highlight it only in the MT output.
- Original sentences and translations have been tokenized, which means there is a space before each punctuation mark. Please, **do not consider these spaces as errors**.

Appraise
Dashboard
esteng0104

0/100 blocks, 7 items left in block
Explainable Quality Estimation: EST - ENG

Tundub , et sellel kaalutlusel on eriline jõud Jimi juhtumil .

— Source text - Click on the words that explain your assessment of translation quality

It seems that this consideration has a special force in the case of Jim .

— Machine translation - Click on the words that explain your assessment of translation quality

https://et.wikipedia.org/wiki/Utilitarianism:_For_and_Against

— Wikipedia article with the context of the translation

Original meaning is closely preserved, only minor mistakes

— Evaluate the quality of the translation using the scale from Completely Incorrect to Perfect Translation.

Reset
Submit

Saving your work

While you are working on the example you can click on the “reset” button if you need to correct your annotation. After you finish working on the annotation item, click “submit”. Your work will be saved and you will access the next annotation item. **Note that after clicking “submit” the annotation cannot be revised.**

Score: 17

In this example, a few keywords are translated correctly, however the original meaning is completely lost due to a mistranslation error ("global rural population") and wrong word order ("end of peace and hostilities" instead of "peace and end of hostilities") arising from a syntactic ambiguity in the source sentence.

(3) Source: Pronksiajal võeti kasutusele pronksist tööriistad , ent käepidemed valmistati ikka puidust

MT: Bronking tools were introduced during the long term , but handholds were still made up of wood .

PE: Bronze tools were introduced during the Bronze Age , but handles were still made of wood .

Context: [Minose kultuur – Vikipeedia](#)

Score: 48

In this example, translation conveys parts of the original sentence but the overall meaning is very difficult to recover due to various major errors (e.g. "Bronze Age" is translated as "long term"). All of the words that justify the low sentence score must be highlighted.

(4) Source: Ootamatult saabub koju ka Ursula , kes seekord jääb üpris viisakaks .

MT: Ursula , too , is coming home in anticipation of being quite polite this time .

PE: Unexpectedly , Ursula comes home as well , but this time she remains quite polite .

Context: ["Õnne 13" jagude loend – Vikipeedia](#)

Score: 60

In this example, the MT output conveys the overall meaning of the source sentence, but contains a major error ("in anticipation of") that makes it sound very strange in English.

It is not obvious which word in the source sentence corresponds to this error. In such cases, make your best guess by aligning source and target words. In this example, it appears that "in anticipation of" stands in place of the pronoun "kes" and we highlight this word accordingly.

Also, note that we highlight the word "ootamatult" ("unexpectedly") in the source sentence, as it is missing from the translation.

(5) Source: Tundub , et sellel kaalutlusel on eriline jõud Jimi juhtumil .

MT: It seems that this consideration has a special force in the case of Jim .

PE: It seems that this consideration has a special weight in Jim 's case .

Context: [Utilitarianism: For and Against – Vikipeedia](#)

Score: 89

In this example, MT output closely preserves the meaning of the source, but is not rendered perfectly in English. We need to highlight minor errors in this case, as they are the ones that justify a high, but imperfect score.

(6) *Source:* Mõnel juhul võib see printsiip tunduda liiga nõudlik .

MT: In some cases , this principle may seem too demanding

PE: In some cases , this principle may seem too demanding

Context: [Practical Ethics – Wikipedia](#)

Score: 96

If the translation is considered perfect or near perfect, no words need to be highlighted.