How STAR Transit NXT can help translators measure and increase their MT post-editing efficiency

Julian Hamm
STAR Deutschland GmbH
Umberto-Nobile-Straße 19
71063 Sindelfingen, Germany
julian.hamm@star-group.net

Judith Klein
STAR Group
Wiesholz 35
8262 Ramsen, Switzerland
Judith.klein@star-group.net

Abstract

As machine translation (MT) is being more tightly integrated into modern CAT-based translation workflows, measuring and increasing MT efficiency has become one of the main concerns of LSPs and companies trying to optimise their processes in terms of quality and performance. When it comes to measuring MT efficiency, STAR's CAT tool Transit NXT offers post-editing distance (PED)[1] and MT error categorisation as two core features of Transit's comprehensive QA module. With DeepL glossary integration and MT confidence scores, translators will also have access to two new features which can help them increase their MT post-editing efficiency.

1 STAR Transit NXT

In the context of today's technology-shaped localisation business, Transit NXT keeps evolving as a sophisticated CAT tool to respond to the needs of language professionals. It does so by supporting a variety of MT providers ranging from STAR's proprietary MT system, commercially available third-party MT providers, through to customisable MT solutions, while also providing the appropriate tools to evaluate MT-based projects and enhance the overall MT post-editing experience.

2 Quality rating

With the *proofreading mode* enabled, translators can mark the MT output in the target language, select an error category and choose a weighting for it. The error categories are based on a slightly condensed interpretation of the SAE J2450[2] metric. The weighting is divided into "minor" and "serious", depending on the impact the translation error might have on the understanding of or action related to the translation. In the Transit NXT quality report, users can set filters to get an overview of the error category distribution within an entire project or individual files and the possible impact these errors may have on the post-editing process. Despite the fact that newer QA metrics such as MQM have been around for quite a while now, our experience has shown that the SAE J2450 metric does provide sufficient coverage of all error categories needed to evaluate the quality of the raw MT output.

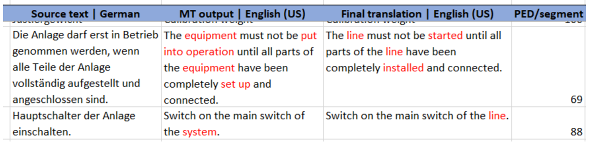
3 Post-editing distance

The *revision mode* allows translators to keep track of all changes made during the editing process. Changes are saved on a per-segment basis and can be conveniently displayed in the segment information window in the Transit NXT editor. In the quality report, users are then provided with an overview of all MT segments. For each MT segment, the report shows the source text, the MT output as well as the final translation, complemented by a per-segment PED value. Changes made during the editing process are highlighted in both the MT output and the final translation column. This new reporting feature is currently being reviewed and

[1] The PED uses a slightly modified Levensthein distance formula. Through string comparison, the PED metric returns a value ranging from 0% (worst) to 100% (best) based on the number of manipulations (additions, deletions, substitutions) in comparison to the total number of characters of a text string.

[2] Standardised metric established by the Society of Automotive Engineers (SAE) for the evaluation of translation quality.

will be released for Transit NXT users in the near future within the Service Pack 15 update cycle[3].



Source text | German	MT output | English (US)	Final translation | English (US)	PED/segment
Die Anlage darf erst in Betrieb genommen werden, wenn alle Teile der Anlage vollständig aufgestellt und angeschlossen sind.	The equipment must not be put into operation until all parts of the equipment have been completely set up and connected.	The line must not be started until all parts of the line have been completely installed and connected.	69
Hauptschalter der Anlage einschalten.	Switch on the main switch of the system.	Switch on the main switch of the line.	88

Figure 1: STAR Transit NXT PED report

While the PED alone can provide valuable insights regarding the reduction in typing actions, it does not consider the cognitive load and actual time spent on the task. However, when being monitored statistically over a longer period of time, it can enable translators to see an increase in efficiency over the course of time, e.g. when switching to a more suitable MT system.

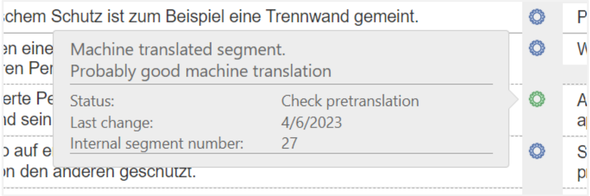
4 Quality report

The Transit NXT *quality report* is a module designed to consolidate the results of all relevant QA checks in Transit and have them readily available in a single report document. The QA report is divided into different main error categories. Each category provides distinct and valuable information, e.g. user-defined *protected strings* missing for translation, error categories and severity for *quality rating J2450*, or the preferred term from the project dictionary for *incorrect terminology*. Each error or inconsistency is accurately logged with the file name, segment number, source language and target language segment content to make it very easy for translators to evaluate and correct the MT output.

5 New smart features for post-editing

For translators, Transit NXT already features a plethora of options to enhance the post-editing experience. The Internal Repetitions (IR) mode helps them identify and correct identical segments that were not translated consistently by MT to avoid unwanted variants. TM validation for MT segments compares the MT output to a highly similar TM translation and visualises the differences between both versions for convenient editing.

A new feature introduced with the latest Transit NXT update is the integration of DeepL glossaries that allows translators to upload a stripped-down copy of their Transit NXT project terminology for supported language combinations and have the preferred terms applied directly to the MT output. This reduces the overall effort needed to correct terminology errors in the MT output and provides greater consistency. As a complement to this, Transit NXT's built-in terminology checks add an extra layer of convenience. Term recognition allows users to visually distinguish whether a preferred term from the project terminology was used in the MT output or not. The second feature to be released in 2023 is referred to as the *MT Confidence Score*[4], which is based on a combination of the modern COMET[5] metric and proprietary AI algorithms. Before editing, translators can visually distinguish between MT suggestions that require a higher or lower level of attention.



chem Schutz ist zum Beispiel eine Trennwand gemeint.

Machine translated segment.
Probably good machine translation

Status: Check pretranslation
Last change: 4/6/2023
Internal segment number: 27

n den anderen geschutzt.

Figure 2: MT confidence score showing an estimation of good (green) MT output

6 Future challenges

Based on our own experience, we are firmly convinced that using the evaluation methods and smart features mentioned above enables us to get a better understanding of the benefits and shortcomings of MT systems in real-world scenarios. Moreover, we argue that even though automated MTQE metrics have become much more reliable over time, evaluation of real-world MT projects still provides us with better insights into the augmented translation experience. However, analysing these parameters is often time-consuming. That said, we see a definite need for developing an advanced PED model that does not only measure edit distance as such but puts it into the bigger context by applying weighting coefficients – based on cognitive load – to the existing error categories. Our first step in developing such a model will be the implementation of AI to automatically classify the changes made during the editing process.

[3] Subject to changes. Screenshot does not show official release version.

[4] The feature is already available in STAR's online editor CLM WebEdit, as part of the CLM workflow.

[5] As seen in: https://virtual.2020.emnlp.org/paper_main.835.html