Analyzing and Predicting MT Utility and Post-Editing Productivity in Enterprise-scale Translation Projects

Alon Lavie
Safaba Translation Solutions
5804 Forbes Avenue, Suite 200
Pittsburgh, PA 15217
alon.lavie@safaba.com

Olga Beregovaya
Welocalize
1820 Gateway Drive, Suite 330
San Mateo, CA 94404
olga.beregovaya@welocalize.com

Michael Denkowski
Safaba Translation Solutions
5804 Forbes Avenue, Suite 200
Pittsburgh, PA 15217
mdenkows@safaba.com

David Clarke
Welocalize
Block G, Cherrywood Business Park,
Loughlinstown, Dublin 18, Ireland
david.clarke@welocalize.com

Abstract

Welocalize has established an MT-driven program for translation and localization services, which is currently deployed for several of its major enterprise clients. At the core of this program are enterprise-optimized Machine Translation engines which are developed and deployed by Safaba Translation Solutions. While the integration of MT and MT post-editing into the translation process results in significant gains in translator productivity and overall project execution velocity, these gains often vary greatly across projects and within projects. Identifying and analyzing the main factors that impact MT utility and post-editing productivity at fine-levels of granularity is thus a critical first step in predicting and improving the expected effectiveness of the MT-based translation process in live enterprise-scale translation projects.

This "user" presentation will focus on the findings of an extensive analysis performed by Welocalize and Safaba on live, enterprise-scale project environments in which MT-based translation processes have been deployed. The data underlying this analysis is based on actual MT post-editing productivity information that was collected on a per-segment basis via a recognized, full-featured open-source CAT tool. The analysis contrasts and correlates the collected segment-level productivity measures with several established MT quality evaluation metrics, human evaluation of a subset of segments by trained post-editors and detailed characteristic properties of the source text. The data is also used to develop segment-level automated quality estimation scores, which can be used to predict the expected utility of MT generated translation segments in future production projects.

Welocalize's objective is to establish a three-dimensional matrix of measures, which can reveal correlations between productivity, expected MT quality and intrinsic properties of the text being translated. Such correlation will allow for more accurate prediction of MT engine performance and expected post-editing productivity for a variety of different source text characteristics.

The sample data selection for the analysis was based on highest/mid/lowest required post-editing time ranges for sentences of the same or similar length.

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 305–307.

We examine a wide range of identifiable source text features, including specific content type categories (i.e. marketing/UI/UA); length of the source segment; source segment morpho-syntactic complexity; presence/absence of predefined glossary terms or multi-word glossary elements, UI elements, numeric variables, product lists, 'do-not-translate' and transliteration lists; as well as certain metadata attributes and their representation in localization industry standard formats ("tags").

The presence and placement of such metadata tags are historically considered to be a major challenge for both MT and post-editors, hence the first step was to analyze the impact that the presence and ratio of the standard XLIFF tags have on the post-editing task duration and factor this impact in the post-editing effort evaluation. A new variable was introduced - a 'tag density ratio' (tags per word) for the machine-translated segments. We analyze the impact of the "tag density ratio" on the overall post-edit time and also its impact on the number of edit visits as compared to 'un-tagged' strings of similar word count ranges. Using string length (word count) ranges, tag quantification, tag density and visit frequency data, several different relationships between these variables are visualized and interpreted. For instance, we test the hypothesis that segments with high tag density exhibit considerably higher than expected post-edit time as compared with low tag density segments of the same length, even if no tagging adjustment is necessary during post-editing. Using a method of calculating tag count and therefore tag density (tags/word) for each individual string from MySQL data exports, we can now identify segments with and without tags, where the translatable content did not require post-editing, and test the hypothesis that tag density results in higher post-editing effort.

The next step in the analysis was to identify segments that contain glossary terms, "DoNotTranslate" elements, URL strings or other identifiable entities and to analyze their post-edit session duration in comparison with segments of similar length with no identified terminology or other "easy-to-manipulate" or "no need to handle" components which inherently simplify the post-editing effort. While this information is not explicitly analyzed via the productivity desktop workbench parser, the event information is captured in the database in raw XML event action form and can be easily extracted and interpreted.

The final step in the study was to perform a morpho-syntactic analysis of the input source sentences and cross-compare this analysis with the pre-defined taxonomy of errors in the machine translation output, that, based on our translators' reports, cause major productivity losses in post editing.

While the current study was performed under the assumption that the post-edited content quality standards are not different from those applied to the traditional "human translation", we have also performed extensive human analysis of the errors found in the post-edited segments and proposed "relaxing" of certain post-editing quality criteria which can lead to additional potential productivity gains.

By analyzing the correlation levels within the three-dimensional matrix of measures, Welocalize has been able to identify and characterize the most advantageous scenarios for MT post-editing, which promise the highest productivity gains, as well as lower productivity gain scenarios, which still result in productivity gains over translation of new words "from scratch".

The same database of three-dimensional matrix of measures was used by Safaba to develop segment-level confidence estimation classifiers, which can then be used to predict the expected quality and utility of MT-generated translations on a segment-by-segment level or at the document level, using the same MT systems

in future production projects. We use two quantifiable measures directly extractable from the collected data as representative of post-editing effort: actual post-edit times recorded for each post-edited segment; and an edit distance measure (TER) of the number of edit operations performed in transforming the MT-generated translation into the final post-edited version of the same segment. A collection of features are extracted and/or calculated for each MT segment. These include both source and target language features, and focus primarily on "extrinsic" features – features that do not require any information that is internal to the MT system itself. The selection of features was guided by the results of the WMT-2012 shared task on quality estimation, which identified a collection of features that were found to be most effective for predicting post-editing effort. The collection of extracted features for each MT segment and their corresponding measures of post-editing effort are then used to train a SVM-based regression classifier for predicting post-editing effort. One of the main challenges in this specific setup was the uneven distribution of segments with different amounts of required post-editing. The available data contains high proportions of segments that require relatively little post-editing effort. While this is indicative of good MT performance overall, it poses significant challenges on accurately identifying "bad" segments which require higher levels of post-editing.

The comprehensiveness of the underlying data set allows us to develop quality estimation classifiers on a large number of languages. We analyze the effectiveness of this approach to quality estimation classification across the set of different languages. To our knowledge, this is the broadest study of quality estimation to-date in terms of language diversity, where the underlying data used in the study was all generated using the same MT technology and collected post-editing effort information.

The presentation will also cover "quality compromise" scenarios, where productivity gains can be achieved via appropriate accuracy/fluency requirements relaxation and/or adjustment of the accuracy/fluency ratio, relative to the specific usage/purpose and intended audience of the content type.