Source Content Analysis and Training Data Selection Impact on an MT-driven Program Design

Olga Beregovaya, David Landan
<first.last>@welocalize.com
Welocalize, Inc.

Description

Clients requiring translation and localization services have come to require an ever-increasing volume of data to be processed, and an unprecedented diversity in the nature of the data to be translated. To meet the increasing demand for translation and the various requirements to the quality of the target output, nearly all language service providers (LSPs) have added machine translation (MT) and various levels of post editing (PE) as integral components of their service offerings.

It has been repeatedly shown that statistical MT engines trained on clean and relevant in-domain data lead to better quality of machine translation output, by using just one of the quality measurement metrics. The importance of corpus preparation and curation and matching the training corpus to the specifics of the content to be translated cannot be overstated. Because of the rapid growth of the amount of data that must be processed, it is imperative that LSPs replace human source content and training corpora evaluations, which are costly both in terms of time and money spent, with a range of programmatic methodologies, which allow for predicting the quality of machine-translated output when specific training data is used, selecting the most suitable translation and post-editing approach and assembling the right workforce for the task.

We employ a large and still-growing suite of tools (both proprietary and through joint academic partnerships) for selecting the best suited dataset matched to the source content to be translated, and estimating the quality of the machine-translated output and the subsequent post-editing effort. To that end, we present several ways that we are working towards automating training data selection and matching it with the source content using a suite of source content analysis tools including:

- Candidate Scorer – a proprietary tool; uses part of speech (POS) n-grams to identify hard-to-translate segments, using a pre-selected corpus that is known to give the worst results, based on human ranking of such segments and post-editing time and distance.
- Source Content Profiler (alpha) – an Industry Partnership CNGL project; uses several features to classify documents into profiles and flags challenges for both machine and human translation
- Perplexity Evaluator – a proprietary tool; generates a matrix of perplexity scores for candidate and control documents against various language models (LMs) built from pre-selected corpora for good and bad results and one custom LM built from historical client in-domain data
- TMTPrime – an Industry Partnership CNGL project; provides a mechanism for automating selection between multiple MT engines, based on source input, using in-domain training data.
- StyleScorer (alpha) – a proprietary tool; scores and ranks candidate source documents according to established style guidelines. In training document selection, StyleScorer learns from a monolingual client corpus that adheres to a desired style, then combines scores from several NLP-based algorithms to generate a final score between 0 and 4 (with 4 being best match to established style).

It has become evident to us that the tools originally created specifically for a single task of either target data selection or source content profiling are often beneficial for both tasks. We present details of the above tools in conjunction with case studies that highlight where each tool has led to improved MT output and/or reductions in post-editing effort. We also present support tools that, while not strictly related to content analysis and data selection, make the outcome of the aforementioned tools and processes easier to interpret.