Streamlining Translation Workflows with StyleScorer

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

Description

The need for quick-turnaround, high-volume machine translation (MT) projects continues to grow in the localization industry. There is a wide range of quality requirements not only across different clients, but often within a single client across different content types (sales & marketing materials, user-generated content, website content, user manuals, etc.). Most clients have style guides or manuals which translators and post-editors are instructed to adhere to, and there may be different styles for the different content types. To help balance the increase in project complexity with clients' needs for faster turnaround times, we created the StyelScorer tool.

StyleScorer compares a new (candidate) document against two or more other documents (the training set); it assigns the candidate document a score between 0 and 4 (higher scores indicate greater stylistic similarity between the candidate document and the training set). StyleScorer generates this score via a weighted combination of several components, including document dissimilarity, perplexity, and unary classification using both neural networks and support vector machines. The candidate document and training set must be written in the same language (for best results, they should be the same locale as well), and the documents in the training set should have internal consistency of style.

We have found StyleScorer to be useful in various stages of the translation workflow by using it on both source- and target-language documents. Many clients wishing to start a new MT program don't have sufficient bilingual assets to train a targeted MT system. By looking for open-source bilingual data where the source-language text is a close stylistic match to the client's training set, we increase the amount of bilingual training data available to build relevant in-domain MT engines.

Once an MT system is deployed, we can use StyleScorer on source-language documents to obtain an estimate of MT output quality by scoring candidate documents against a training set created from the MT engine training set. We can then use StyleScorer on a target-language training set to estimate the amount of post-editing effort required to bring the MT output in line with the desired target-language style. The benefits here are two-fold: documents above a given threshold can be automatically marked as passing, and post-editors can focus their attention on the lower-scoring documents.

We are now also experimenting with using StyleScorer as part of the linguistic QA process. Randomly selected post-edited documents are checked against the target-language test set. Low scoring documents are given a second round of review, with two possible outcomes: further post-edits are required before the document is given a passing grade, or the low-scoring document is deemed acceptable. In the latter case, we update the training set with new target-language documents to accurately capture the acceptable style patterns.