Dataset Statement and Disclaimer:

This dataset was created for the purpose of training an automatic method for estimating the quality of machine translation output at run-time, without relying on reference translations. We created this dataset from Russian Reddit forums (75%) and Russian WikiQuotes (25%). We then translated the text to English using a pre-trained FairSeq model, and translators rated the quality of the translations using a direct assessment score. We included Reddit data since colloquial text is a challenge for machine translation, and Russian proverbs from WikiQuotes to test machine translation on short sentences with unconventional grammar. These sources were chosen because they are farther outside the training data used for machine translation, which is typically more formal text as in Wikipedia or news. Translation accuracy is relatively high for text that is close to training data, so we chose data that would include a wider range of quality scores.

This content reflects the full spectrum of Russian vernacular and colloquial usage, both past and present, which is essential to create effective ML models. Since, for many applications (like hate-speech or toxic comment detection), users must be able to recognize potentially objectionable content, or text that is inadvertently translated to such. As a result, this dataset could include offensive content. Further, because this data-set includes full paragraphs, posts, and conversations between Reddit users, we chose not to filter individual sentences, since this would reduce the utility of the data-set for document-level quality estimation, another important technical challenge for MT.

Although we has taken reasonable steps to verify that hyperlinks in the dataset do not lead to malicious content, the dataset still may still include links to inappropriate content. We in no way endorses nor condones such words, views, or activities.