Abstract

Machine translation has become ever more prominent in society with Google translate acknowledging over 500 million users in 2016. With this reliance on machine translation, it is important to determine quality and whether human parity has been reached. Automatic metrics like BLEU scores can be used to estimate the quality of machine translations but these methods are often considered unreliable. Therefore, manual evaluations campaigns are carried out including at the annual workshop on statistical machine translation (WMT). There has been an appeal in the machine translation community to start evaluating at the document-level to give annotators more context. As such, the direct assessment method used at WMT was adapted to suit a document-level evaluation instead of isolated sentences and segments. However, the method has fallen short at providing full document context while also failing to give annotators a chance to edit previous segments.

The research undertaken aimed at addressing this problem of providing full document-level context while also ensuring high-levels of quality control for the crowd-sourced assessments. This project is the first to explore the use of post-editing at the document-level for machine translation with crowd-sourced annotations. The benefits of this approach over existing evaluation methodologies are that it gives full document context, removes subjective numerical judgements and it can specify where MT systems go wrong. Evaluations were carried out by researchers as well as crowd-sourced workers. In addition, inter-annotator agreements levels were calculated against these post-edits using Cohen's kappa.

The results demonstrated some correlation in terms of machine translation system rankings with official results that were obtained at WMT 20. In the crowd-sourced evaluation campaign, 2 cluster groups of rankings emerged compared to 4 cluster groups of rankings from official WMT results. In terms of inter-annotator agreement for post-editing, it was found that annotators showed moderate agreement when choosing to edit or not edit a segment ($\kappa = 0.4821$) but this agreement was not as strong when crowd-sourcing ($\kappa = 0.2826$). Furthermore, the QC mechanisms proved effective and allowed for a manual inspection of submitted data, something that current machine translation evaluations can not feasibly do. Overall, the work carried out demonstrated that crowd-sourcing document-level post-edits is a feasible approach to machine translation evaluation.