Toward an Automated, Task-Based MT Evaluation Strategy

John S. White

Litton PRC

There has recently been a convergence, like an astrological conjunction, of needs and capabilities in language handling activities, driven mostly from external trends. Our expectation of access to information has, consciously or not, transcended our awareness of language barriers. We expect somehow, unless we know better, that online information exists which can be automatically manipulated in such a way as to tell us what we need to know regardless of such trivialities as source language.

Of course, we do know better. We know that there are no perfect, transparent MT systems, and further that there are not a body of systems that cover all the possible languages from which we desire information. New language pairs must be developed, and existing ones improved, far more quickly than we yet know how.

Meanwhile, there is at last an awareness of a context in which MT is to occur, namely the flow of a variety of combinations that make up end-to-end information handling processes. This vision requires that MT, and indeed all text/language components, be as human-independent as absolutely possible, and the humans remaining in the loop have no special expertise beyond the subject matter handled in the information process.

So the convergence is one of greater automation, greater speed in development, and much greater sensitivity to the context in which MT operates. But while these trends are happening, the evaluation of MT remains very labor intensive, very slow to develop, and largely standalone.

The reasons why MT evaluation cannot be automated in a simple way are well known: since there are many correct translations of any expression, there is no single ground truth. And since determining which translation is "better" than another (even if both are good or both are poor) is a matter of linguistic, and often subjective, judgments, the best MT evaluation techniques must avail themselves of human intuition, requiring in turn a large enough sample size, abundant human factors controls, and so on. Finally, such evaluations may generally tell us something about the potential of a particular approach to translation as an end in itself, but not directly about how useful a system may be, in more or less its present state, within some end-to-end information process.

In recent months a US government effort has been underway to develop a methodology for MT evaluation that is directly germane to task-based metrics. Like other evaluation approaches, it too is human judgment-intensive, drawing from expert information specialists. But this method holds the promise of capturing these judgments in a test corpus that can be run automatically by MT systems.

This paper describes the process for distilling these human judgments into test patterns. It address the issues associated with creating such a test set, and the potential of transcending related problems such as the for need test corpora for every language pair and direction among the many hundreds of languages of interest. The paper suggests some means for generating these test sets so that we can have evaluation approaches with the same properties as the future MT systems we hope to evaluate: automatic, rapidly developed, and relevant to their intended tasks.