We present results from a new machine translation comprehension test, similar to those developed in previous work (Jones et al., 2007). This test has documents in four conditions: (1) original English documents; (2) human translations of the documents into Arabic; conditions (3) and (4) are machine translations of the Arabic documents into English from two different MT systems. We created two forms of the test: Form A has the original English documents and output from the two Arabic-to-English MT systems. Form B has English, Arabic, and one of the MT system outputs. We administered the comprehension test to three subject types recruited in the greater Boston area: (1) native English speakers with no Arabic skills, (2) Arabic language learners, and (3) Native Arabic speakers who also have English language skills. There were 36 native English speakers, 13 Arabic learners, and 11 native Arabic speakers with English skills. Subjects needed an average of 3.8 hours to complete the test, which had 191 questions and 59 documents. Native English speakers with no Arabic skills saw Form A. Arabic learners and native Arabic speakers saw form B.

The overall comprehension results for English natives reading Form A were 88% for the original English texts, 53% for MT1, and 76% for MT2. System level BLEU scores were 0.0624 and 0.2158 respectively. Comprehension scores were not strongly correlated with BLEU scores. For the Arabic language learners who saw Form B, the comprehension results were 91% for English, 46% for Arabic, and 76% for MT2. For the Arabic native speakers who saw Form B, comprehension results were 82% for English, 80% for Arabic, and 72% for MT2. The Arabic learners, who had an average of 2.5 semesters of Arabic instruction at the college level, demonstrated comprehension at a level between that of MT1 and MT2 as read by native English speakers. No MT results were as good as native speakers reading their native language.

We used the standardized language skill descriptions defined by the Interagency Language Roundtable (ILR); see (ILR, 2014). To measure machine translation capabilities, as opposed to human foreign language capabilities, we constructed a variant of the Defense Language Proficiency Test, following the general DLPT design principles, but modified to measure the quality of machine translation. This test is a multiple-choice format, ILR-based machine translation test format described in the paper “ILR-Based MT Comprehension Test with Multi-Level Questions” by Jones et al. in the proceedings of HLT 2007. Test documents were rated for ILR reading skills and were split between Levels 2, 2+ and 3. Questions were also rated for ILR level: Level 1, 2, 2+, and 3; comprehension results generally reflected the difficulty levels.

References

This work is sponsored by the Defense Language Institute under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.