Abstract

We present results from two new experiments in which educated English-native speakers answered questions from a machine translated version of Arabic language tests based on the Interagency Language Roundtable (ILR) skill levels. We compare current state-of-the-art machine translation (MT) results with professional reference translations as a baseline. The first experiment uses a test based directly on the Defense Language Proficiency Test (DLPT) design, containing passages at Level 2(L2) and Level 3(L3), and was administered to 35 native English speakers and took approximately six hours for each subject to complete. There were three primary test conditions: baseline high-quality human translations; audio MT from recorded Arabic speech and text MT from Arabic text passages. The overall comprehension results were as follows: Text MT: 73%@L2; 57%@L3; Audio MT: 67%@L2; 47%@L3; Human translation: 91%@L2; 87%@L3. The conventional passing threshold for the DLPT is 70%. The second experiment uses a modified test design in which lower level questions are presented within more difficult text passages, i.e., questions at Level 1 to Level 3 are asked within the context of a Level 3 passage. These test materials were drawn from the DARPA/GALE 2006 Arabic dry-run evaluation data for news broadcasts (BC), newsgroups (NG), broadcast news (BN) and broadcast conversation (BC). This test was administered to 90 native English speakers and took approximately five hours for each subject to complete. Preliminary scores indicate that the comprehension rate for the high quality human translations ranged from 87% to 94% (all levels and genres). The breakdown by level for MT was: 50%@L1; 68%@L1+; 74%@L2; 76%@L2+; 63%@L3. By genre: BC@37%; BN@69%; NG@76%; NW@78%. The lower performance at L3 is to be expected; the lower L1 and L1+ performance is likely due to mistranslated names. We present an analysis as well as some intriguing misinterpretations for garbled machine translation passages.